NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

"Translational Engineering in Health and Medicine"

Application of Machine Learning methods to predict critical Multiple Myeloma events

Postgraduate Diploma Thesis

Marina I. Georgoula

Supervisor: George Matsopoulos, Professor

Athens, June 2024

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

"Translational Engineering in Health and Medicine"

Application of Machine Learning methods to predict critical Multiple Myeloma events

Postgraduate Diploma Thesis

Marina I. Georgoula

Supervisor: George Matsopoulos, Professor

The postgraduate diploma thesis has been approved by the examination committee on 11th June 2024

1st member	2nd member	3rd member
George Matsopoulos	Dimitrios Fotiadis	Leonidas Alexopoulos
Professor	Professor	Associate Professor
School of Electrical and Computer Engineering, NTUA	Dept. of Materials Science and Engineering University of Ioannina	School of Mechanical Engineering, NTUA

Athens, June 2024

……..

Marina I. Georgoula
Graduate of the Interdisciplinary Postgraduate Programme,
“Translational Engineering in Health and Medicine”,
Master of Science,
School of Electrical and Computer Engineering,
National Technical University of Athens

Abstract

Multiple Myeloma (MM), a hematological malignancy, presents a complex challenge due to its multifaceted nature driven by genetic, molecular, and clinical factors. This thesis aimed to leverage comprehensive multi-source data from the Multiple Myeloma Research Foundation CoMMpass study to develop accurate and reliable predictive models addressing three critical events in MM. The case studies examined include prediction of relapse, treatment response, and mortality risk. This research constructed predictive models employing longitudinal clinical, phenotype, and transcriptomic data, to anticipate relapse at the first and fifth year since diagnosis and treatment initiation, since a key characteristic of MM is frequent relapses after initial therapeutic responses, signaling drug-resistant clones or disease progression. The models demonstrated robust predictive performance, enabling timely identification of patients at risk for relapse. These forecasting models could allow clinicians to implement preemptive strategies, such as salvage therapies or intensified treatment regimens. Furthermore, predicting treatment response is crucial in MM for guiding therapeutic decisions, optimizing dosages, and customizing surveillance strategies. This task was divided into three subtasks reflecting different contexts of treatment response. For the first and fifth year of first-line treatment, models were trained as binary classifiers using baseline cross-sectional data and longitudinal panel data as two different subtasks. The XGBoost model excelled at both prediction horizons with baseline input data, while the Attention-LSTM model effectively handled sequential data, suggesting further optimization. Additionally, a third subtask predicted response to any line therapy six months ahead, using six months of patient history, with LSTM-variant models yielding high performance results. Significant predictors for treatment response were also identified including treatment-related information, status of MM disease at diagnosis as well as some laboratory measurements. Lastly, this thesis integrated clinical and transcriptomic features to develop prognostic models for 5-year survival in MM, stratifying patients into distinct risk cohorts. The models exhibited strong predictive performance, aiding shared decision-making, resource allocation, and tailoring palliative care. Clustering transcriptomic data provided a grouped risk stratification tool, with SHAP analysis revealing significant predictors, including disease status at diagnosis and duration of treatment responsiveness. In summary, this thesis highlights the potential of advanced predictive modeling in enhancing MM patient management, offering actionable insights for relapse prediction, treatment response, and long-term survival, thereby overall improving quality of life and patient outcomes.

Keywords

Multiple Myeloma, CoMMpass study, Predictive models, Machine learning, Deep learning, Relapse prediction, Treatment response, Mortality risk stratification, Longitudinal data, Patient management.

Περίληψη

Το Πολλαπλούν Μυέλωμα (ΠΜ) ως πλασματοκυτταρική κακοήθεια, αποτελεί μια σύνθετη πρόκληση λόγω της πολυπλοκότητάς του, που προέρχεται από γενετικούς, μοριακούς και κλινικούς παράγοντες. Αυτή η διπλωματική εργασία στοχεύει στην αξιοποίηση ολοκληρωμένων πολυδιάστατων δεδομένων από το Multiple Myeloma Research Foundation CoMMpass study για την ανάπτυξη αξιόπιστων προγνωστικών μοντέλων που εξετάζουν τρία κρίσιμα γεγονότα που σχετίζονται με την ασθένεια. Οι περιπτώσεις που εξετάστηκαν περιλαμβάνουν την πρόβλεψη υποτροπής, την ανταπόκριση στη θεραπεία και τον κίνδυνο θνησιμότητας. Συγκεκριμένα, κατασκευάστηκαν προγνωστικά μοντέλα χρησιμοποιώντας διαχρονικά κλινικά, δημογραφικά και δεδομένα γονιδιακής έκφρασης, για να εκτιμηθεί η υποτροπή κατά το πρώτο και πέμπτο έτος από τη διάγνωση και την έναρξη της θεραπείας, δεδομένου ότι ένα βασικό χαρακτηριστικό του ΠΜ είναι οι συχνές υποτροπές, υποδεικνύοντας ανθεκτικούς κλώνους ή εξέλιξη της νόσου. Τα μοντέλα επέδειξαν ισχυρή προγνωστική απόδοση και στους δύο ορίζοντες πρόβλεψης, επιτρέποντας την έγκαιρη αναγνώριση ασθενών σε κίνδυνο υποτροπής. Επιπλέον, η πρόβλεψη της ανταπόκρισης στη θεραπεία είναι κρίσιμη στο ΠΜ για την καθοδήγηση των θεραπευτικών αποφάσεων, τη βελτιστοποίηση των δόσεων και την εξατομίκευση των στρατηγικών παρακολούθησης. Σε αυτήν την περίπτωση, υλοποιήθηκαν τρία υπο-προβλήματα που αντικατοπτρίζουν διαφορετικά πλαίσια ανταπόκρισης στη θεραπεία. Αρχικά, τα μοντέλα εκπαιδεύτηκαν ως δυαδικοί ταξινομητές χρησιμοποιώντας δεδομένα που συλλέχθηκαν στην πρώτη επίσκεψη και διαχρονικά δεδομένα, προβλέποντας την ανταπόκριση στο πρώτο και πέμπτο έτος της πρώτης γραμμής θεραπείας. Το μοντέλο XGBoost διακρίθηκε με υψηλή απόδοση, έχοντας δεδομένα εισόδου βασικής γραμμής, ενώ το μοντέλο Attention-LSTM χειρίστηκε αποτελεσματικά τα διαδοχικά δεδομένα, υποδεικνύοντας την ανάγκη για περαιτέρω βελτιστοποίηση. Επιπλέον, στο τρίτο υπο-πρόβλημα, τα μοντέλα παραλλαγής του LSTM μοντέλου, προέβλεψαν με υψηλή ακρίβεια και ευαισθησία την ανταπόκριση σε οποιαδήποτε γραμμή θεραπείας έξι μήνες νωρίτερα, χρησιμοποιώντας έξι μήνες ιστορικού ασθενούς. Επίσης, εντοπίστηκαν σημαντικοί προγνωστικοί παράγοντες για την ανταπόκριση στη θεραπεία, συμπεριλαμβανομένων των πληροφοριών που σχετίζονται με τη θεραπεία, την κατάσταση της νόσου κατά τη διάγνωση και ορισμένων εργαστηριακών μετρήσεων. Τέλος, αυτή η εργασία ενσωμάτωσε κλινικά και χαρακτηριστικά γονιδιακής έκφρασης για την ανάπτυξη αποδοτικών προγνωστικών μοντέλων σχετικά με την 5ετή επιβίωση, κατατάσσοντας τους ασθενείς σε διακριτές ομάδες κινδύνου. Συνοψίζοντας, αυτή η εργασία αναδεικνύει τις δυνατότητες των προηγμένων προγνωστικών μοντέλων στη βελτίωση της διαχείρισης των ασθενών με ΠΜ, βελτιώνοντας έτσι τη συνολική ποιότητα ζωής και τα αποτελέσματα των ασθενών.

Λέξεις Κλειδιά

Πολλαπλούν Μυέλωμα, Μελέτη CoMMpass, Προγνωστικά μοντέλα, Μηχανική μάθηση, Βαθιά μάθηση, Πρόβλεψη υποτροπής, Απόκριση στη θεραπεία, Διαστρωμάτωση κινδύνου θνησιμότητας, Διαχρονικά δεδομένα, Διαχείριση ασθενών.

Acknowledgments

I would like to express my deepest gratitude to all those who have supported me throughout the completion of this thesis. First and foremost, I am profoundly grateful to my advisors, Professor George Matsopoulos as my supervisor, as well as Professor Dimitris Fotiadis, for their invaluable guidance, insightful feedback, and unwavering support.

I extend my heartfelt thanks to the faculty and staff of TEAM Master of Science program at NTUA for providing a stimulating academic environment and for their assistance in various stages of my MSc journey.

This research would not have been possible without the data and resources provided by the Multiple Myeloma Research Foundation CoMMpass study. These data were generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org and www.themmrf.org).

Lastly, I owe an immense debt of gratitude to my family for their belief, support and encouragement.

Thank you all for your support and contributions to this work.

Table of Contents

Table of Tables

Table of Figures

Abbreviations

AI	Artificial Intelligence
ASCT	Autologous hematopoietic Stem Cell Transplantation
AUC	Area Under the Curve
AUC-PR	Area Under the Precision-Recall Curve
BiLSTM	Bidirectional Long Short-term Memory network
BUN	Blood Urea Nitrogen
CT	Computer Tomography
CNN	Convolutional Neural Network
DL	Deep Learning
EORTC	European Organization for Research and Treatment of Cancer
FL	Focal Lesion
FPR	False Positive Rate
GEO	Gene Expression Omnibus
GEP	Gene Expression Profile
HR	Hazard Ratio
IgA	Immunoglobulin A
IgG	Immunoglobulin G
IgM	Immunoglobulin M
ISS	International Staging System
iFISH	Interphase Fluorescent In Situ Hybridization
IMiDs	Immunomodulatory Drugs
LDH	Lactate Dehydrogenase
LR	Logistic Regression
LSTM	Long Short-term Memory network
MM	Multiple Myeloma
ML	Machine Learning
MLP	Multilayer Perception

MMRF	Multiple Myeloma Research Foundation
MRI	Magnetic Resonance Imaging
OS	Overall Survival
PCA	Principal Component Analysis
PET-CT	Positron Emission Tomography-Computed tomography
PI	Proteasome inhibitor
QoL	Quality of Life
RF	Random Forest
R-ISS	Revised International Staging System
RMSProp	Root Mean Square Propagation
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic Curve
RR	Ridge Regression
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TPM	Transcripts per Million
TPR	True Positive Rate
WBC	White Blood Cell
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

Introduction

Background and Motivation

Multiple myeloma (MM) remains a challenging hematological malignancy characterized by clonal proliferation of plasma cells within the bone marrow, leading to debilitating skeletal destruction, renal impairment, and immunodeficiency. Multiple Myeloma has an age-standardized incidence rate of 1.8% [1] and an estimated number of 188,000 cases worldwide (2022) [2], while it was estimated that in the future 2045 there will be an increase of 321,000 new cases [3]. Specifically in the US, the estimated number of deaths in 2024 is 12,540, with a 5-year survival (2014-2020) up to 61%. Typically, the diagnosis begins with a thorough clinical evaluation, including a detailed medical history and physical examination. Laboratory tests play a crucial role in MM diagnosis, with serum and urine protein electrophoresis, serum immunofixation, and serum-free light chain assays aiding in the detection of monoclonal proteins characteristic of the disease. Imaging studies such as skeletal surveys, Magnetic Resonance Imaging (MRI), and Positron Emission Tomography-Computed tomography (PET-CT) scans are employed to assess bone involvement, detect lytic lesions, and evaluate disease extent. Additionally, bone marrow examination, including bone marrow aspiration and biopsy, remains a cornerstone in confirming MM diagnosis, assessing plasma cell infiltration, and determining disease stage. Patients with MM often exhibit common but non-specific clinical manifestations, including fatigue, acute kidney injury, hypercalcemia, normocytic anemia, immunosuppression, weight loss, and bone pain. The advent of these symptoms frequently results in delays in diagnosis compared to other malignancies, serving as the main indications for diagnosis. These indications are defined as the CRAB criteria specifically including a) hypercalcemia greater than 2.75 mmol/L or serum calcium greater than 0.25 mmol/L(Calcium), b) creatinine greater than 177 μmol/L (Renal), c) Hemoglobin less than 10 g/dL or a decrease of 2 g/dL (Anemia), d) at least one lytic bone lesion on X-ray, CT, or PET scan (Bone) [4]. The updated treatment criteria from the International Myeloma Working Group (IMWG) incorporate the traditional CRAB criteria (Calcium, Renal, Anemia, Bone) along with three supplementary criteria for identifying active disease [5] incorporating a) Marrow plasmacytosis of 60% or higher, b) a ratio of involved to uninvolved serum light chains of 100 or higher and c) the existence of at least two focal bone lesions, each exceeding 5 mm in size on MRI. In 2015, the International Myeloma Working Group introduced the Revised International Staging System (R-ISS) for Multiple Myeloma to improve patient stratification for personalized treatment recommendations based on prognosis. The R-ISS [6] incorporates chromosomal abnormalities detected via interphase Fluorescent In Situ Hybridization (iFISH) and serum Lactate Dehydrogenase (LDH) levels alongside the traditional International Staging System (ISS) criteria. Patients are classified into three stages: Stage I, Stage II, and Stage III. Remarkably, the five-year overall survival rates for R-ISS stages I, II, and III were reported as 82%, 62%, and 40%, respectively. Treatment response is typically monitored through serum or urine paraprotein measurements, or serum free light chain assessment if the ratio of involved to uninvolved free light chains is abnormal. For the minority of patients with non-secretory MM lacking these measurable serum biomarkers, plasma cell percentage and skeletal imaging, particularly MRI or PET/CT, are utilized for monitoring [7].

Despite advancements in treatment modalities over recent decades, MM continues to pose significant clinical management hurdles due to its inherent heterogeneity, diverse clinical manifestations, and variable treatment responses among patients. The treatment landscape for this malignancy encompasses a variety of modalities aimed at controlling disease progression, managing symptoms, and improving patient outcomes [4]. First-line therapy often includes combination regimens incorporating Immunomodulatory Drugs (IMiDs) such as lenalidomide or thalidomide, Proteasome Inhibitors (PI) like bortezomib or carfilzomib, and corticosteroids such as dexamethasone. Stem cell transplantation, particularly Autologous Hematopoietic Stem Cell Transplantation (ASCT), may be considered for eligible patients to achieve deep and durable responses. Maintenance therapy with IMiDs or monoclonal antibodies, such as daratumumab or elotuzumab, may be employed to prolong remission and delay disease progression. Additionally, targeted therapies directed against specific molecular pathways, novel agents undergoing clinical trials, and supportive care measures, including bisphosphonates and erythropoiesis-stimulating agents, play crucial roles in the comprehensive management of MM. The standard induction regimen in many countries remains VTD (bortezomib, thalidomide, and dexamethasone), which has shown superiority over conventional chemotherapy. The management of MM relies heavily on accurate risk stratification and timely intervention to optimize therapeutic decisions and improve patient outcomes. Traditional prognostic factors such as disease stage, cytogenetic abnormalities, and serum biomarkers have provided valuable insights into disease progression and treatment response. However, the inherent complexity and heterogeneity of MM pose challenges in accurately predicting critical events, including disease progression, relapse, and overall survival.

The challenges encountered in diseases like Multiple Myeloma have prompted the rise of precision medicine, a method of managing diseases that considers an individual's genetic profile, environment, and lifestyle. This enables doctors and researchers for optimized matching of patients with effective treatments. Given the intricate and varied nature of MM, the success of treatment relies on tailoring approaches specific to each patient. The emergence of precision medicine aligns with an increasing tendency of patients engaging in clinical decision-making. Enhancing patients' comprehension of their unique disease characteristics and possible disease trajectory provides knowledge positively influencing their outcomes, thereby reinforcing the role of precision medicine in the future.

The advent of Artificial Intelligence (AI) and especially Machine Learning (ML) methodologies has revolutionized the landscape of biomedical research and clinical practice by enabling the extraction of meaningful insights from vast and complex datasets. In the domain of oncology, ML techniques hold immense promise for unraveling the intricacies of MM pathogenesis, refining risk stratification strategies, and facilitating informed clinical decision-making. By leveraging advanced computational algorithms, ML frameworks offer unprecedented opportunities to integrate multi-dimensional data sources encompassing genomic profiles, clinical parameters, imaging studies, and treatment histories, thereby fostering a comprehensive understanding of MM biology and progression, as well as a holistic approach to risk assessment and outcome prediction.

The motivation for this research stems from the pressing clinical need to develop accurate and reliable predictive models for critical events in MM, including treatment response, mortality, and relapse. Predictive modeling offers a promising avenue for leveraging longitudinal clinical

data, genomic profiles, and treatment response metrics to identify high-risk patients, optimize treatment strategies, and inform personalized care plans. By integrating advanced machine learning techniques with comprehensive clinical and molecular data, this research aims to fill existing gaps in predictive modeling and potentially to enhance decision-making clinical practices by enabling early intervention, guiding treatment decisions, and improving long-term outcomes for MM patients.

Objectives

The primary objective of this study is to develop and validate predictive models for the accurate and timely prediction of critical events in Multiple Myeloma, including relapse, treatment response and mortality. By systematically reviewing the state-of-the-art AI & ML applications in predicting critical events in MM, this study aims to:

1. Provide a comprehensive overview of existing methodologies, datasets, and predictive features utilized in MM prediction studies.
2. Evaluate the performance and clinical utility of ML models in prognosticating disease progression, treatment response, and overall survival.
3. Identify key challenges and limitations in current AI approaches and propose avenues for future research and clinical implementation.
4. Ultimately, contribute to the growing body of evidence supporting the integration of AI in precision medicine initiatives for MM patients, with the ultimate goal of improving outcomes and quality of life.

In summary, by addressing these interconnected objectives, this thesis aspires to harness the transformative potential of AI & ML methodologies to empower clinicians with actionable insights for optimizing therapeutic outcomes, minimizing disease-related morbidity, and improving overall survival in this challenging hematological malignancy.

Problem Formulation

The complex interplay of genetic, molecular, and clinical factors underlying MM pathogenesis underscores the need for innovative approaches. In this context, predictive modeling offers a promising avenue for leveraging comprehensive clinical and molecular data to develop accurate and reliable models for predicting critical events in MM. Based on the State-of-the-Art, the specific objectives driving this research, providing a framework for the development and validation of predictive models to address the unmet clinical needs in MM management, include:

1. **Relapse prediction:** Despite initial responses to therapy, MM frequently exhibits a propensity for disease relapse, often indicating the emergence of drug-resistant clones or disease progression. By harnessing longitudinal clinical data and incorporating dynamic biomarkers, this thesis seeks to construct predictive models for anticipating disease relapse

at specific prediction horizons (‘Year 1’ and ‘Year 5’), and preemptively identifying patients at heightened risk for clinical deterioration. Timely recognition of impending relapse could empower clinicians to implement preemptive salvage therapies, enrollment in clinical trials, or intensification strategies aimed at forestalling disease progression and preserving quality of life.

2. **Treatment Response prediction**: A key determinant of patient outcomes in MM, treatment response encompasses a spectrum ranging from stringent complete response to progressive disease. By harnessing AI & ML algorithms, this thesis aims to develop advanced predictive models capable of discerning the likelihood of achieving favorable treatment responses based on patient phenotypes and clinical profiles. Such models hold the potential to guide therapeutic selection, dose optimization, and surveillance strategies tailored to individual patient profiles, thereby enhancing treatment efficacy, and minimizing adverse events. Specifically, three subtasks are formulated in this case. Initially, focusing on first-line treatment, which remains critical for MM disease trajectory, its treatment response will be predicted at specific prediction horizons (‘Year 1’ and ‘Year 5’), both with cross-sectional data collected at baseline, as well as longitudinally. Moreover, regarding all forms of treatments, responsiveness will be predicted sequentially in time, providing dynamic monitoring, real-time insights and adaptation, enabling clinicians to make timely adjustments based on the most recent data. This optimization of patient outcomes also helps in minimizing adverse effects.
3. **Mortality Risk prediction**: Prognostication of overall survival remains a cornerstone of clinical decision-making in MM, guiding treatment intensity, supportive care measures, and end-of-life planning. Through integration of clinical and transcriptomic features, this thesis endeavors to develop prognostic models capable of stratifying patients into distinct risk cohorts with varying survival probabilities. For this objective, two approaches will be implemented. Firstly, mortality risk will be predicted within 5-years since patient enrollment, utilizing data collected at baseline. For the second approach, clustering will initially be applied on transcriptomic data to extract patient clusters, hence each cluster’s data will be used to train the model to predict cluster-based mortality risk.

Related Work

State of the Art

Multiple Myeloma presents a complex and heterogeneous clinical landscape, necessitating accurate prognostic tools to guide treatment decisions and improve patient outcomes. In recent years, AI & ML techniques have emerged as promising tools for predicting critical events of malignancies including disease progression, treatment response [8],[9], and overall survival, outperforming conventional statistical methods.

Accurate prediction of treatment response is essential for optimizing therapeutic strategies and improving patient outcomes. Studies have employed a variety of ML algorithms to predict

treatment response across different pathologies, ranging from traditional supervised learning methods to more complex DL architectures. Conventional ML models such as Support Vector Machines (SVMs), Random Forest (RF), and Logistic Regression (LR) are frequently employed algorithms for binary classification tasks, distinguishing responders from non-responders to particular therapies [10], [11], [12]. More recently, Deep Learning methods including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have also been investigated for analyzing high-dimensional omics data and imaging studies to forecast treatment outcomes [13].

Addressing these objectives in Multiple Myeloma, despite significant advancements in treatment options due to new drug discoveries [14], [15], [16] many MM patients experience relapse, primarily due to large interindividual differences in treatment response caused by gene expression changes associated with chemoresistance [17], [18], [19]. Understanding these alterations is crucial for evaluating the effectiveness of targeted treatments and choosing appropriate therapies promptly. However, despite technical progress in high-throughput drug screening, analyzing the resulting data remains challenging. Although differential Gene Expression Profile (GEP) analyses have identified GEP as a useful predictive marker for MM risk stratification, none of these profiles have been drug-specific so far [20], [21]. For instance, while PIs are commonly used in MM treatment, response to therapy varies among patients, underscoring the need to identify precise predictive markers [22]. Recent studies have utilized machine learning algorithms to predict treatment response based on gene expression patterns. The utilization of RF and RF-survival models demonstrated promising results in discriminating subjects based on their response to PIs in Mitra et al. study [23]. Borisov et al. [24] conducted an experiment involving 53 MM patients with data from the Gene Expression Omnibus database (GEO) (https://www.ncbi.nlm.nih.gov/geo/), treated with two different protocols: bortezomib, doxorubicin, dexamethasone (PAD) and bortezomib, cyclophosphamide, dexamethasone (VCD). RNA sequencing profiles of good and poor responders were analyzed, identifying five genes upregulated in good responders, namely GRB14, MAF, FGFR3, IGHA2, and IGHV1-69. Using RF, linear SVM and Ridge Regression (RR), binomial Naive Bayers, and Multilayer Perceptron (MLP) classifiers, good and poor responders within each treatment group were distinguished, while integrating data-trimming with the FloWPS system enhanced the classifiers effectiveness. The MLP model accomplished an area under the curve (AUC) of 89%, however the classifiers couldn't transfer across different patient groups likely due to treatment protocol variations and MM heterogeneity. In a recent study, Povoa et al. [25] introduced a novel approach called Multi Learning Training (MuLT), which integrates supervised, unsupervised, and self-supervised learning methodologies. MuLT was designed to assess the prognostic significance of diverse treatment outcomes in MM. By analyzing clinical and gene expression levels data from 1525 MMRF CoMMpass study patients, it was revealed that gene expression profiles contribute to therapy sensitivity prognosis and encapsulate genetic alterations identified through Fluorescence in-situ hybridization (FISH) analysis. Through cross-validation experiments, MuLT demonstrated a therapy sensitivity prediction accuracy of 68.7% and AUC of 61.54%. Importantly, the findings indicated that approximately 17.07% of included MM patients could potentially benefit from alternative chemotherapy regimens as first-line treatment options. Upon investigation of specific treatment schemas and their efficacy, Ubels et al. [26] introduced Simulated Treatment learning

signatures (STLsig) as a new method for predicting therapeutic benefits in MM patients upon diagnosis, potentially assisting in optimal treatment selection. This method led to the identification of gene profiles associated with enhanced survival outcomes from proteasome inhibitor (PI) therapy compared to other treatments. Within a cohort of 910 MM patients, STLsig pinpointed two gene complexes predictive of favorable responses to the PI bortezomib. In the "benefit" group, bortezomib exhibited a Hazard Ratio (HR) of 0.47, contrasting with an HR of 0.91 among the "no benefit" group.

Promising survival estimators have also emerged from the fusion of clinical and laboratory parameters with GEP. Focusing on a cohort of 15 patients that have relapsed to prior treatments, Paulus et al. [27] presented a Computational Biology Modeling (CBM) tool that integrates various genetic alterations to predict key molecular pathways utilized by MM cells for survival. By simulating MM cell architecture, it forecasts drug responses and resistance mechanisms, paving the way for more personalized disease management strategies. Utilizing 50 variables, including age, ISS stage, serum β2-microglobulin level, first-line therapy type, and the expression of 46 genes, Mosquera Orgueira et al. [28] employed ML-based personalized prediction to assess overall survival (OS) across six first-line treatments for MMRF CoMMpass study patients, employing Random Forest model that achieved a C-index of 81.8%. Forecasting survival models were also developed by Ren et al. [29], where decision tree analysis for risk stratification in one-year OS prediction achieved an AUC of 68%, including demographics, ISS stage, low or high Ubiquitin Proteasome Pathway Risk Score (UPPRS), single-cell RNA-sequencing. Park et al. [30], designed an ML-assisted methodology for personalized predictions on the response and overall survival of 514 newly diagnosed MM patients from the Catholic Research Network for MM database (CARE-MM) using bortezomib-lenalidomide-dexamethasone (VRD) or lenalidomide plus dexamethasone (RD) as their first-line therapy. Including six input variables, namely diabetes, LDH, serum kappa, creatinine, *t*(4;14), and *t*(11;14), the XGBoost model to predict OS in frontline VMP-treated patients, achieved a AUC of 75% and 71% accuracy on the test cohort. Similarly, another XGBoost model for predicting the OS of RD-treated patients achieved an AUC of 93% and accuracy of 78%, based another set of six covariates, namely ISS stage, *t*(4;14), kappa, creatinine, LDH, and M protein level. Using these models, risk stratification was performed for each group of patients, highlighting that these ML survival models for treatment selection could lead to improved overall survival. Exploiting two randomized clinical trials, Cook et al. [31] developed an overall survival risk profile for MM patients unqualified for stem-cell transplantation, utilizing multivariate Cox regression model with least absolute shrinkage and selection operator penalty term, including WHO performance status, International Staging System, age, and C-reactive protein concentration as predictors.

Although certain patients experience prolonged remission with initial triple combinations containing immunomodulatory drugs and proteasome inhibitors, combined with high-dose treatment plus autologous stem cell transplantation, prompt remission has emerged as an independent risk factor for resistance to subsequent treatments and reduced overall survival. To this end, Kubasch et al. [32] implemented ML algorithms utilizing input data from 563 MMRF CoMMpass study patients, such as sociodemographic, clinical, and cytogenetic data, to predict early relapse of MM patients. After applying the Boruta algorithm for feature selection, the four most important features selected were the best treatment response recorded within the first year after diagnosis, followed by Autologous stem cell transplant during the first year of diagnosis, patient age and beta-2-microglobulin. In this study, the Gradient Boosting Classification model

emerged as the most effective in forecasting early relapse, achieving an accuracy of 73%, correctly classifying 82% of patients with early relapse and 69% of those without in the test dataset. Risk of progression was also modeled by utilizing traditional risk models [33], and assessing the correlation between risk factors, such as FISH, ISS, RISS, GEP, age at progression, minimal residual disease status and focal lesion (FL) assessments with PET-CT, and the likelihood of mortality and disease progression utilizing a multivariable Cox proportional hazards model. This analysis resulted in a strong association of increased number of focal lesions with a worse patient outcome, highlighting the importance of identifying focal lesions, as well as validation of positive impact of minimal residual disease negative status after first relapse on progression-free survival and overall survival.

In summary, recent advancements in biomedical data analysis, powered by machine learning and deep learning, offer promising avenues for understanding and managing Multiple Myeloma. However, challenges such as data variability and ethical considerations persist. Despite this, the integration of AI holds significant potential to streamline clinical workflows and enhance patient care. Moving forward, careful navigation of AI utilization alongside adherence to ethical principles is crucial for maximizing its impact on MM research and patient outcomes.

Materials & Methods

Data Source

The dataset employed in this thesis is part of the Multiple Myeloma Research Foundation CoMMpass study (themmrf.org/finding-a-cure/personalized-treatment-approaches/#study). This study is a prospective observational longitudinal genomic-clinical study (NCT01454297) of more than 1,100 patients recruited since 2011. The CoMMpass study dataset is considered one of the most extensive Myeloma publicly available datasets, while aiming to exploit both genomic and clinical landscape of MM disease for patient profiling, risk stratification and exploration of new findings. Patient enrollment to the study is defined by both inclusion and exclusion criteria. Specifically, inclusion criteria consist of:

- Patient is at least 18 years old.
- Patient has been diagnosed with symptomatic MM with measurable disease that includes at least one of the following: Serum M protein ≥ 1g/dl Urine M protein ≥ 200 mg/24 hours Involved free light chain level ≥ 10 mg/dl and an abnormal serum free light chain ratio (< 0.26 or > 1.65)
- The patient is a candidate for systemic therapy that includes an IMiD® (e.g., lenalidomide, pomalidomide, thalidomide) and/or proteasome inhibitor (e.g., bortezomib, carfilzomib) as part of the initial regimen.
- No more than 30 days from baseline bone marrow evaluation as per this protocol to initiation of first-line therapy.
- Patient has read, understood and signed informed consent.

On the other hand, exclusion criteria are defined as:

- Patient had another malignancy within the last 5 years (except for basal or squamous cell carcinoma, or in situ cancer of the cervix).
- Patient is already receiving systemic therapy for MM (a single dose of bisphosphonates and up to 100 mg total dose of dexamethasone or equivalent corticosteroids are permitted prior to registration on study).
- Patient is enrolled in a blinded clinical trial for the first-line treatment of Multiple Myeloma. Patients may be enrolled in subsequent clinical trials as long as continued access to data and tissue, as per this protocol, is not prohibited.

Furthermore, based on the latest available documentation (themmrf.org/wp-content/uploads/2020/05/MMRF_CoMMpassWP_final.pdf, April 2020), some baseline demographics of 1,143 enrolled participants are described in Table 1. It should be noted that 60% of patients are still enrolled, whereas 25% of patients have died due to disease progression or other reason, based on the latest analysis. Focusing on lines of therapy and treatment regimens, and specifically in the first line of therapy, the predominant regimen utilized was the triplet combination consisting of Velcade®-Revlimid®-Dex. Although Velcade®-Dex and Revlimid®-Dex, both doublet combinations, were frequently prescribed for first-line treatment, triplet combinations are regarded as the standard-of-care regimen for this indication.

Table 1. MMRF CoMMpass Cohort demographics

Baseline Characteristic	**N=1,143**
Age (median (min, max))	63 (27, 93)
Gender - Female (n (%))	453 (40)
Ethnicity (n (%))	
Hispanic/Latino	76 (8)
Non-Hispanic/Non-Latino	861 (89)
Other	34 (3)
Race (n (%))	
White	742 (76)
Black/African American	161 (17)
Asian	18 (2)
American Indian/Alaskan Native	1 (<1)
Other	49 (5)

Data Analysis

The MMRF CoMMpass study has recruited newly diagnosed patients with MM from the United States, Canada, Spain, and Italy. This dataset consists of clinical and phenotype parameters, which are gathered initially and then every three months throughout the eight-year monitoring period. In the context of this thesis, the dataset version IA17 was used, by requesting access to the platform portal from MMRF administrators (https://research.themmrf.org/). The features are further analyzed into sub-categories for better comprehension.

Baseline data

As baseline data, are considered data that were collected at baseline, namely at the first visit of patients accompanied by some overall patient profile information. The features in this category include all demographic variables such as age, gender, ethnicity, weight, height, along with patient profile information such as death, MM status, IgG type, IgA type, ISS disease stage, creatinine levels, date of death, and ECOG performance status. Some distribution details are provided below (Figure 1).

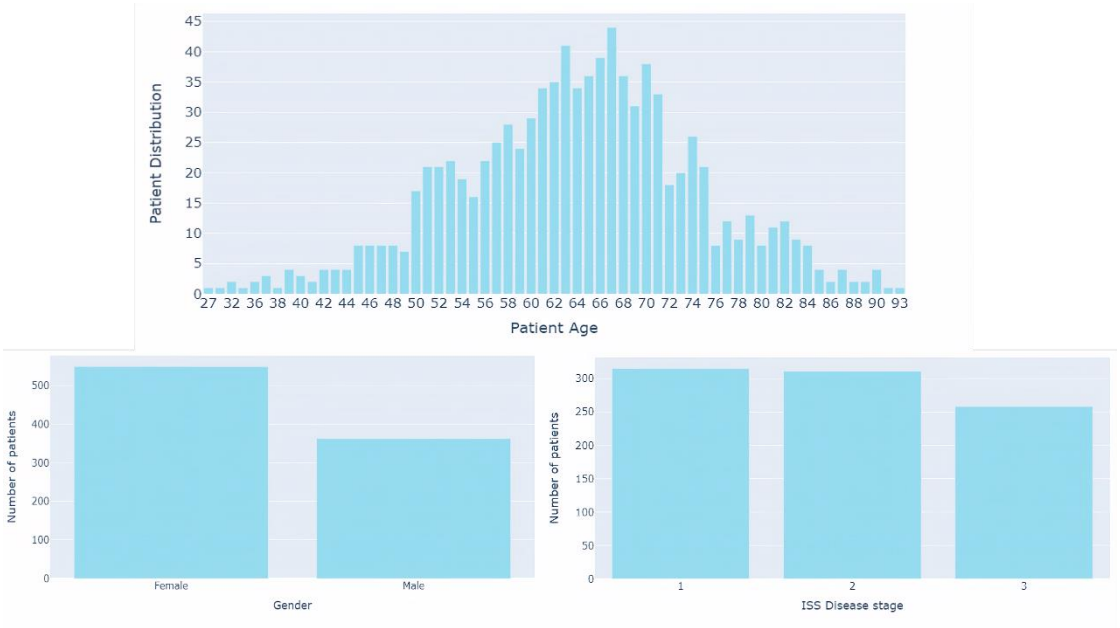

Figure 1. (Top center) Distribution of patient age, (Left) Gender distribution and (Right) ISS Disease distribution.

Treatment data

Also, features regarding treatment are collected across time for each patient, such as treatment name and combination, line of therapy, eligibility and application of stem cell autologous

transplant (Figure 3), number of treatment agents, treatment response, best response for each treatment combination and disease progression. In total, 150 different treatment combination schemas were found in the dataset for first line therapy, with the top 20 visualized in Figure 2. All available treatment options recorded in the cohort treatment planes are listed below (Figure 4), with the combination of Bortezomib-Lenalidomide-Dexamethasone being most applied.

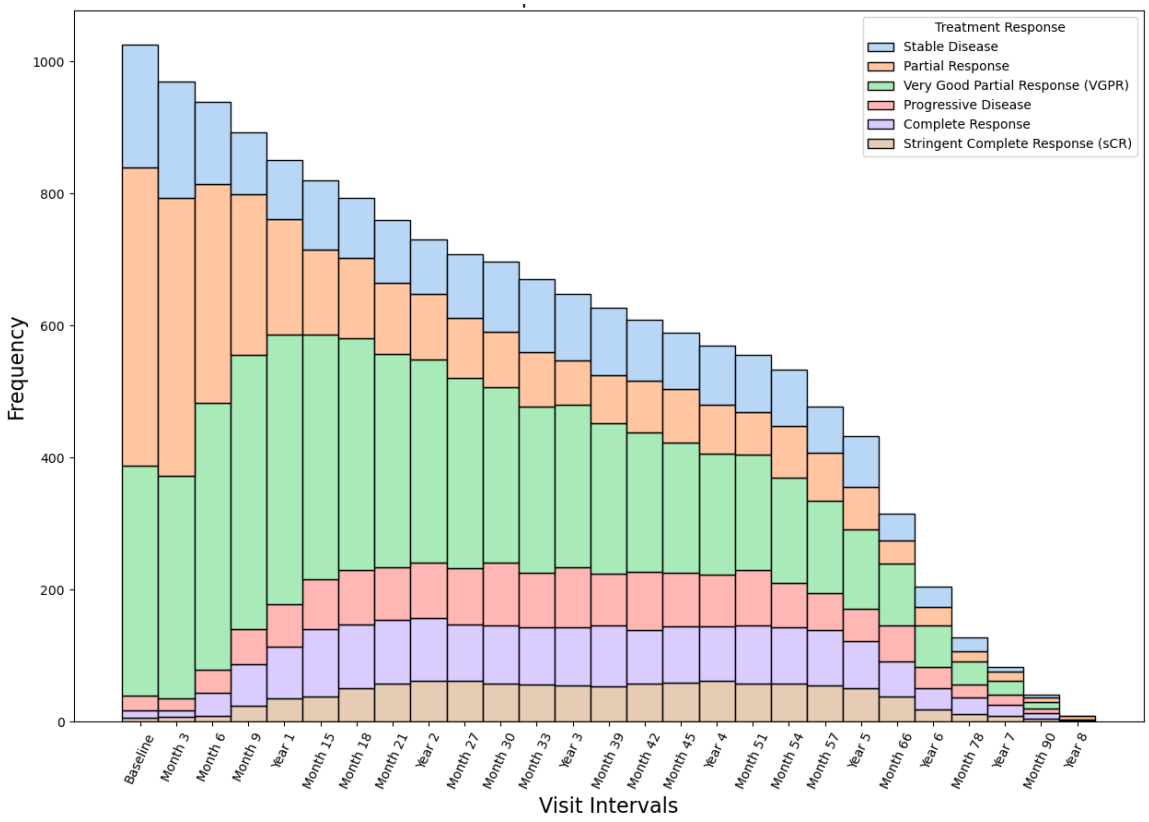

Figure 2. Treatment response distribution across visits.

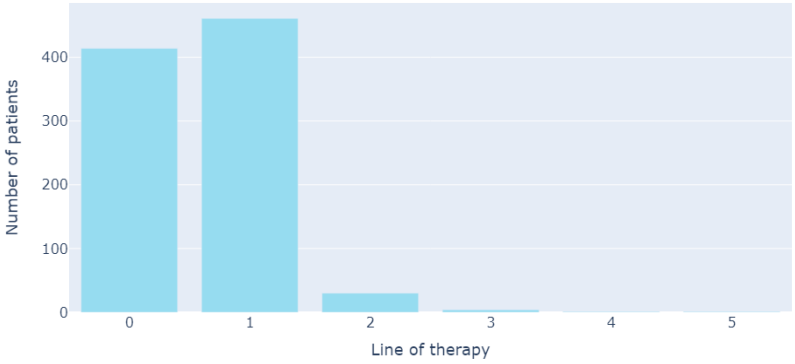

Figure 3. Stem Cell Autologous transplant in each line of therapy.

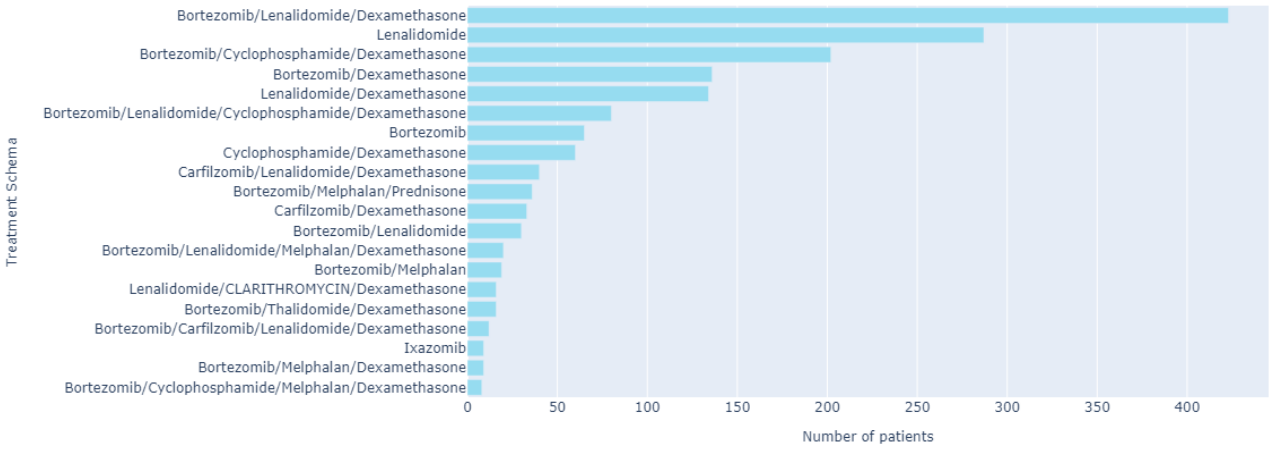

Figure 4. Treatment options in dataset

Lab measurements

Additional lab measurements data are recorded related to patient health state and disease-related state. Specifically, bone assessment values, blood chemistry values (i.e. albumin, BUN, calcium, creatinine, glucose, and total protein) (Figure 6), complete blood counts (i.e. absolute neutrophils, hemoglobin, WBC, and platelets) and serum immunoglobulins (i.e. IgG, IgA, IgM, lambda, kappa, and M-protein) are included (Figure 5). The study protocol also evaluated the physical and cognitive abilities of the patient, the side effects, as well as the social context at each visit through questionnaires.

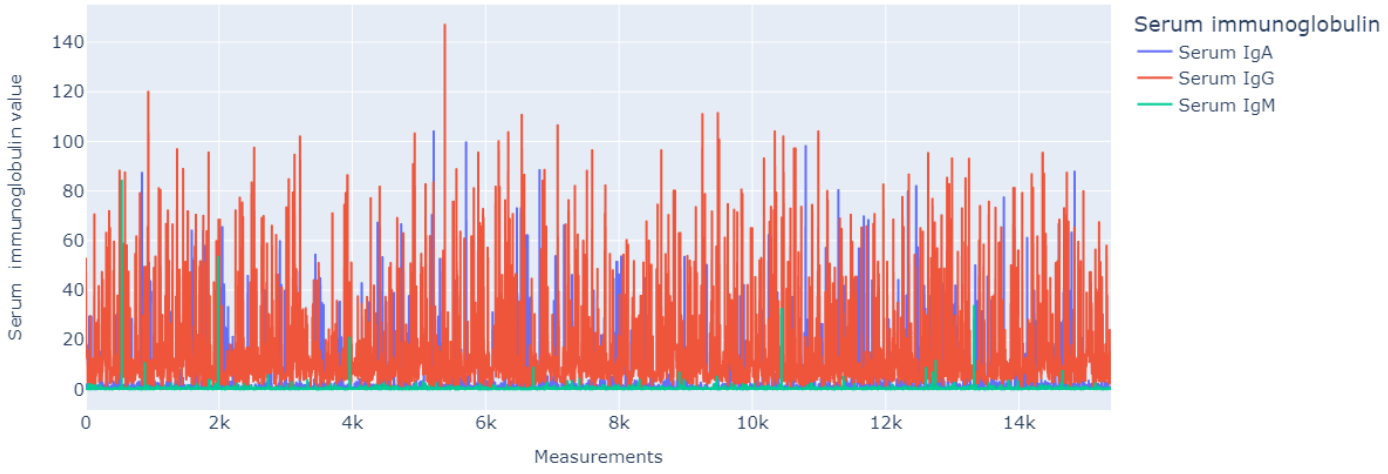

Figure 5. Serum Immunoglobulin values

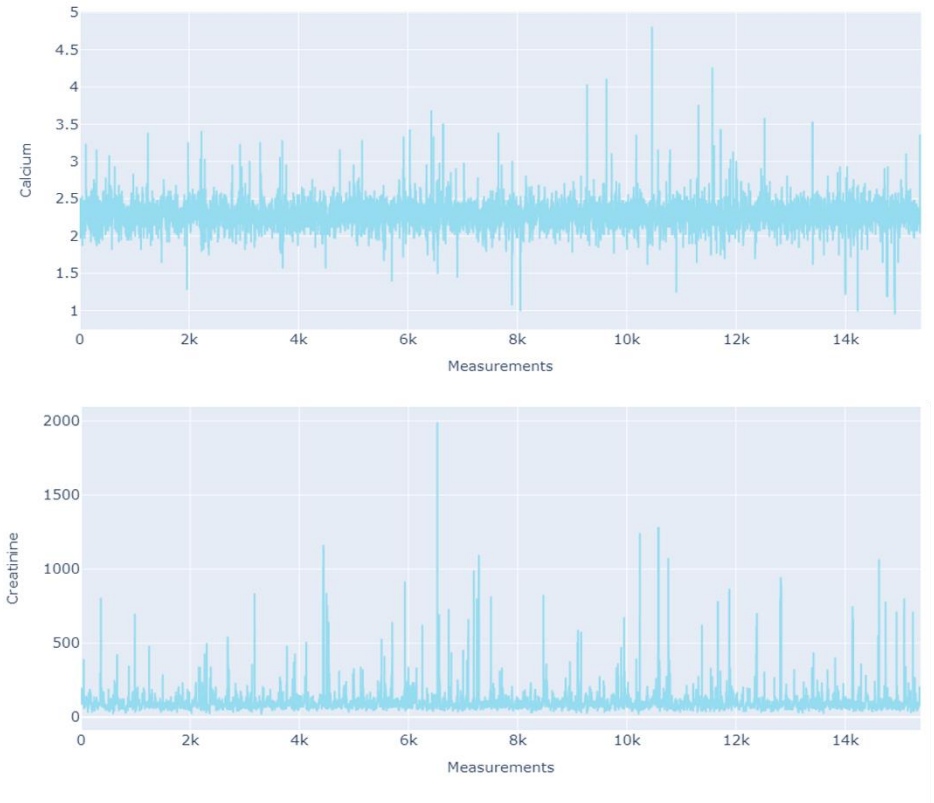

Figure 6. Calcium and Creatinine values across measurements

RNA sequencing

RNA sequencing (RNA-seq) data refers to the comprehensive set of information obtained from sequencing the RNA molecules present in a biological sample. This technology allows for the quantification of gene expression levels, providing insights into which genes are active, to what extent, and at what specific times under various conditions. Gene expressions [34], [35] acts as an "on/off switch" to control when and where RNA molecules and proteins are made and as a "volume control" to determine how much of those products are made. RNA-seq captures a snapshot of the transcriptome, the complete set of RNA transcripts, including mRNA, rRNA, tRNA, and non-coding RNAs. This data contributes to understanding gene regulation, discovering novel transcripts, identifying splicing variants, and studying the dynamic changes in gene expression patterns in response to different environmental or physiological stimuli. In this dataset, tumor specimens are included as a comprehensive characterization through whole-genome sequencing (WGS), whole-exome sequencing (WES), and RNA-seq upon diagnosis and at each subsequent progression event. However, in this thesis, only RNA sequencing data were exploited for further analysis. Specifically, the Salmon method [36] was utilized to map RNA sequencing reads to transcriptome reference sequences to identify differentially expressed genes and pathways.

Questionnaires

Two main questionnaires are included in this dataset, the EORTC QLQ-C30 and the EORTC QLQ-MY20. The European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30 questionnaire is a widely used tool designed to assess the quality of life of cancer patients [37]. Developed by the EORTC, it consists of 30 items that measure a range of physical, emotional, and social health domains, as well as symptoms commonly experienced by cancer patients, such as fatigue, pain, and nausea. The QLQ-C30 is highly regarded for its reliability and validity, making it a standard instrument in both clinical trials and routine oncology practice worldwide. Its comprehensive approach aids healthcare providers in understanding the impact of cancer and its treatment on patients' overall well-being, facilitating more personalized and effective care. The EORTC also developed the Multiple Myeloma module quality of life (QoL) questionnaire (QLQ-MY20) in 1999, and it has been widely utilized alongside the EORTC core quality of life questionnaire (EORTC QLQ-C30) in both clinical trials and everyday practice, to assess the quality of life of patients with MM.

Preprocessing

The MMRF CoMMpass study provides a rich dataset for understanding the molecular and clinical characteristics of MM patients. However, before applying machine learning methods for predicting critical events related to this malignancy, several preprocessing steps are necessary to ensure data quality and compatibility with the chosen models. In this section, the preprocessing steps for MMRF CoMMpass study including 911 patients' data, are outlined. By performing these preprocessing steps, we ensure that the MMRF CoMMpass study data is appropriately prepared for subsequent machine learning analyses, enhancing the robustness, interpretability, and generalization ability of the developed models, ultimately contributing to more accurate and clinically relevant predictions.

Data Cleaning

The first step was to assess the overall data quality and accordingly apply data cleaning methods on 568 original features. Since the data were provided in different files, each data file was examined separately at first, and were merged at a later stage. Three files were examined in total, one file including patient profile information, especially all data collected at the first patient visit, one file including patients' data from each visit, and one file including gene expression estimates data for the majority of patients. Starting from checking each file's duplicate rows per patient, no duplicate entries were found. Moreover, data cleaning encompasses the identification and resolution of inconsistencies or anomalies that could adversely impact the validity and reliability of analyses. In the MMRF CoMMpass study data, this may involve detecting and addressing outliers, which are data points significantly deviating from the majority of observations. Outliers can arise due to measurement errors, data entry mistakes, or genuine anomalies in the underlying phenomena being studied. Techniques such as visual inspection and statistical Z-score method were employed to identify and handle outliers appropriately.

Furthermore, data cleaning procedures extend to the reconciliation of conflicting or redundant information across different variables to ensure coherence and accuracy in the dataset. One such inconsistency case were categorical features with different capitalization of their

categories (ex ‘bright’, ‘Bright’). The most critical aspect of data cleaning involves handling missing values, which are prevalent in real-world datasets like MMRF CoMMpass due to various reasons such as measurement errors or incomplete data collection. Therefore, the identification and handling of missing values was implemented, in order to avoid sparse datasets and patient entries with no information. In case of features missingness over 70% of values, the feature was entirely dropped from the dataset. It is important to note that in this dataset, missing values could be misleading, because in some cases the missing value indicated the absence of a specific procedure or variable. For example, in the case of ‘SS_HYPERCALCEMIA’ categorical feature representing hypercalcemia, the only category present was ‘Checked’, while the missing value stands for ‘Not checked’. In such cases, the feature was converted into a binary variable, where ‘1’ represented the recorded value (i.e. ‘Checked’). After dropping the features with great missingness, this missing values examination was also performed patient-wise. Patients with more than 50% of missing values across all remaining features were removed from the dataset. Furthermore, visit days (‘VISITDY’) which were missing were mapped based on the 3-month interval each visit was assigned to (‘VJ_INTERVAL’), by choosing a random day from this interval. For continuous variables missing, forward fill imputation was applied for each patient, while categorical variables were assigned either a new category for missing values, especially for lab measurements data, marked as ‘No record’, or filled with the corresponding category reflecting neutral or negative input (ex. ‘No’). The final stage of this step was to ensure consistency of feature values format, defining only one data type (float, string) for each feature.

By meticulously addressing issues of missingness, inconsistency, and outliers, data cleaning endeavors to enhance the reliability and trustworthiness of the MMRF CoMMpass study data, empowering researchers to derive meaningful insights and make informed decisions in subsequent analyses and modeling efforts.

Time series alignment & merging

Given the longitudinal nature of the MMRF CoMMpass study dataset, and especially the data file including patient visits data since for each 3-month interval data were collected (starting from Baseline), a critical task involved organizing all the patient data available from different files, into these 3-month intervals. Each of these intervals was considered a different time point, serving as a reference point for each patient visit. Patients with fewer than one-year interval data (i.e. < 5 time points), were not further included in the modeling procedure. This threshold was defined mainly by the prediction tasks which demand a minimum patient data window of five time points. As a result, baseline and visit data were merged into one dataset.

For defining relapse as an outcome, time alignment had also to be performed since the variable providing relapse information (‘D_PT_pddy’) referred to the exact day of relapse from baseline. In this case, the day of relapse was mapped to the corresponding 3-month visit interval it belonged to, in order to reflect relapse on a visit level. Similarly, for representing the mortality outcome, the variable ‘D_PT_deathdy’ referring to the exact day of patient death, was calibrated to the visit-interval axis.

Outcome variables specification

To represent the outcome variables for the prediction tasks, features 'treatment response', 'mortality' and 'relapse' had to be created. Regarding 'treatment response', the original feature 'AT_TREATMENTRESP' was utilized and renamed, which was already a categorical variable including the different response levels at each visit interval. For mortality, as aforementioned, the variable 'D_PT_deathdy' was exploited and transformed into 'mortality', which was initially mapped to visit-interval level. Specifically, mortality was kept only at a yearly basis, namely with a maximum prediction horizon of 5 years, the variables 'mortality_year1', 'mortality_year2', 'mortality_year3', 'mortality_year4' and 'mortality_year5' were created as binary variables. Accordingly, relapse outcome was represented as a binary variable also at a yearly basis creating variables 'relapse_year1', 'relapse_year2', 'relapse_year3', 'relapse_year4' and 'relapse_year5'. To further exploit the outcome variables by line of therapy the integer variable 'line' was utilized.

Feature Selection

Feature selection aimed at identifying and retaining the most relevant variables for this thesis' modeling tasks. In the context of predicting critical Multiple Myeloma events, feature selection involves choosing clinical and phenotype features that are most likely to influence the occurrence and progression of the disease. Clinical features such as patient demographics, laboratory test results, and medical history provide valuable information about the patient's health status and treatment history. However, not all features contribute equally to predictive performance, and some may introduce noise or redundancy into the model. After merging different patient data files, there were features found withholding similar or even identical information. These features were merged or only one was kept (ex. multiple features for treatment schema). In addition, features with only one unique value across all rows and patients were dropped since there is no variability potentially aiding to capture patterns in the modeling procedure (ex. 'enr'). Variables related to the study enrollment and reasons for dropping out of the study were also not included in the dataset. For some modeling tasks, RF Classifier was also employed as feature selector, leveraging the algorithm's inherent ability to assess feature importance during model training. By analyzing the impurity decrease achieved by each feature across multiple decision trees, Random Forest identifies the most informative features for prediction. Features with higher importance scores are considered more influential in predicting the target variable, thereby enabling effective feature selection. This approach not only helps in dimensionality reduction but also enhances model interpretability and generalization by focusing on the most relevant features for the task at hand. In the case of gene expression estimates file, where expression estimates are denoted as transcripts per million (TPM) for each gene, only data at baseline are extracted, whereas values are already normalized per patient ID, but only for 770 patients.

Overall, feature selection techniques aim to identify a subset of informative features while discarding irrelevant or redundant ones, thereby improving model interpretability, generalization, and computational efficiency.

Encoding and Standardization

In machine learning, encoding and standardization are crucial preprocessing steps aimed at preparing data for analysis and modeling [38], [39]. Encoding involves transforming categorical

variables into numerical representations that AI & ML algorithms can understand and process effectively. Categorical variables, such as gender or treatment response, are initially represented as text labels, which need to be converted into numerical values. One common encoding technique is one-hot encoding, where each category is represented by a binary vector, with each dimension indicating the presence or absence of a particular category. This ensures that the model can differentiate between different categories without assigning any ordinal relationship between them. Another method is label encoding, where each category is assigned a unique integer label. Both encoding methods were applied, depending on the context of each categorical variable.

For numerical features, standardization was applied, which focuses on rescaling numerical features to have a mean of 0 and a standard deviation of 1. This process, also known as Z-score normalization, ensures that all features have a similar scale, preventing features with larger magnitudes from dominating those with smaller magnitudes during model training. Standardization is particularly important for algorithms that rely on distance-based metrics, such as k-nearest neighbors and support vector machines. By standardizing the features, these algorithms can treat all features equally, leading to more stable and reliable model performance. Additionally, standardization helps in interpreting the model coefficients or feature importances, as the features are on the same scale, facilitating comparisons between them. Importantly, when applying standard scaling, it's crucial to perform the procedure separately on the training and test datasets. This is because the mean and standard deviation used for scaling should be computed solely based on the training data to prevent data leakage and ensure that the test data remains unseen during preprocessing. By scaling the training and test datasets separately, we maintain the integrity of the test data, enabling accurate evaluation of model performance on unseen samples. This practice reflects real-world scenarios where the model needs to generalize well to new, unseen data.

Class imbalance handling

Class imbalance occurs when the distribution of classes in a dataset is highly skewed, with one class significantly outnumbering the others. This can pose a challenge for machine learning models, as they may become biased towards the majority class and struggle to accurately predict the minority class (e.g., fewer instances of critical MM events compared to non-events). One approach to mitigate this issue is random undersampling [40], where instances from the majority class are randomly removed from the training dataset until a more balanced class distribution is achieved. By reducing the number of instances in the majority class, random undersampling helps prevent the model from being overly influenced by the dominant class and encourages it to learn more discriminative features for the minority class. Specifically for the prediction pipelines implemented in this thesis, random undersampling is performed on a patient-level. However, while random undersampling can be effective in balancing class distributions and improving model performance, it comes with the risk of information loss, as valuable data points may be discarded during the process.

When addressing class imbalance in ML, another method commonly used, especially in tree-based algorithms like XGBoost, is scaling the positive weight of the minority class. Specifically, in XGBoost, the `scale_pos_weight` parameter allows us to assign a higher weight to instances belonging to the minority class compared to those in the majority class during the training

process. By increasing the weight of the positive (minority) class the model focuses more on correctly classifying these instances, thereby reducing the bias towards the majority class. This approach effectively balances the influence of each class on the model's learning process, ensuring that it pays sufficient attention to both classes. However, it's essential to tune the `scale_pos_weight` parameter carefully to avoid overcompensating and causing the model to become overly sensitive to the minority class, which could lead to decreased performance on the majority class or even overfitting.

AI Methodologies

Artificial intelligence techniques have emerged as powerful tools for analyzing complex datasets and extracting meaningful patterns in various domains, including healthcare. In the context of Multiple Myeloma, such algorithms are implemented, both supervised and unsupervised, for this thesis experiments.

Supervised and unsupervised learning are two fundamental paradigms in ML, each serving distinct purposes and addressing different types of problems [41]. In supervised learning, the algorithm learns from labeled data, where each input is associated with an output label or target variable. The goal is to learn a mapping function that can accurately predict the output for new, unseen inputs. This is typically achieved through the use of algorithms such as regression for continuous target variables or classification for categorical target variables.

On the other hand, unsupervised learning deals with unlabeled data, where the algorithm aims to uncover hidden patterns or structures within the data without explicit guidance. Unlike supervised learning, there are no predefined output labels to guide the learning process. Instead, unsupervised learning algorithms seek to identify similarities or differences among data points and group them into clusters or extract meaningful features. Clustering algorithms like K-means and hierarchical clustering are commonly used in unsupervised learning to partition data into clusters based on similarities between data points. Principal Component Analysis (PCA) [42], a dimensionality reduction technique, was also used to visualize high-dimensional data and extract important features.

While supervised learning is suited for tasks where labeled data is available and the goal is to make predictions, unsupervised learning is useful for exploring and understanding the underlying structure of data when labels are absent or difficult to obtain. In many real-world applications, both supervised and unsupervised learning techniques are employed in conjunction to gain a comprehensive understanding of the data and to build more robust models. This unified approach to machine learning allows practitioners to leverage the strengths of each paradigm to tackle complex problems and extract valuable insights from data. In this thesis, supervised learning models are utilized to make predictions for binary or multilabel classification problems, while unsupervised methods are used to implement patient profiling as a first stage, to further develop cluster-based classification models.

Supervised learning

Attention-LSTM

The Attention Long Short-Term Memory (LSTM) model is a neural network architecture designed to enhance the performance of sequential data processing tasks by selectively attending to relevant parts of input sequences [43], [44], [45]. This model builds upon the conventional LSTM architecture (Figure 7), which is known for its ability to capture long-range dependencies and temporal patterns in sequential data.

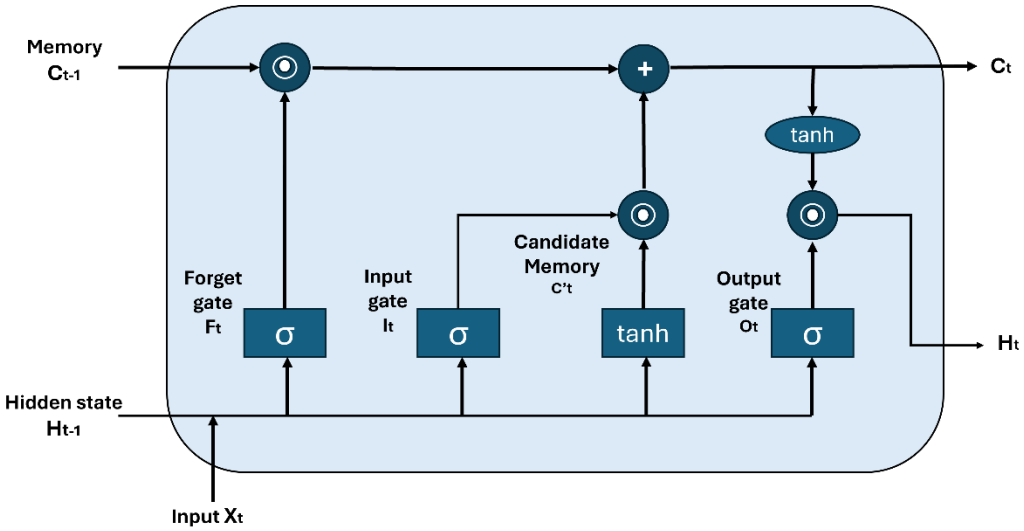

Figure 7. LSTM Architecture

Long Short-Term Memory (LSTM) is a type of RNN architecture designed to address the vanishing gradient problem. LSTMs utilize specialized memory cells with gating mechanisms to selectively retain or forget information over time. Analytically, LSTM model includes:

- Input Gate: Determines how much new information to add to the cell state. It consists of the input gate (I_t) that decides which values from the input to update.

$$I_t = \sigma(W_i \cdot [H_{t-1}, x_t] + b_i) \quad (1)$$

- Forget Gate: Determines how much of the previous cell state (C_{t-1}) to retain or forget. It takes both the current input (X_t) and the previous hidden state (H_{t-1}) as input and outputs a value between 0 and 1, indicating the proportion of each cell state element to retain.

$$F_t = \sigma(W_f \cdot [H, X_t] + b_f) \quad (2)$$

- Cell State Update: The cell state C_t is updated by combining the information retained from the previous state with the new candidate values (C'_t), weighted by the input and forget gates.

$$C'_t = tanh\,(W_c \cdot [H_{t-1}, X_t] + b_c) \quad \textbf{(3)}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot C'_t \quad \textbf{(4)}$$

- Output Gate: The output gate (O_t) regulates how much of the cell state is exposed to the next hidden state. It considers both the current input and the previous hidden state and generates the next hidden state (H_t) based on the updated cell state.

$$O_t = \sigma(W_o \cdot [H_{t-1}, X_t] + b_o) \quad \textbf{(5)}$$

$$H_t = O_t \odot tanh\,(C_t) \quad \textbf{(6)}$$

Where:

- X_t is the input at time step t.
- H_{t-1} is the previous hidden state.
- C_{t-1} is the previous cell state.
- I_t, F_t, O_t, C'_t are the input gate, forget gate, output gate, and candidate memory cell activation vectors, respectively.
- W and b are the weight matrices and bias vectors to be learned during model training.
- σ is the sigmoid function.
- $tanh$ is the hyperbolic tangent function.

By utilizing these gates and memory cells, LSTM networks can effectively learn and retain information over long sequences, making them well-suited for various sequential data tasks such as natural language processing, time series prediction, and speech recognition.

At its core, the Attention LSTM model incorporates an attention mechanism, which dynamically weights the importance of different elements within an input sequence [46]. This attention mechanism enables the model to focus more on informative segments of the input sequence while suppressing irrelevant or redundant information, thereby improving the model's ability to extract meaningful features from the data. The attention mechanism operates through a series of steps during the computation of each output step in the LSTM layer. Initially, the model computes a set of attention weights (Equation 7) by evaluating a compatibility function between the current hidden state of the LSTM layer and the hidden states of all input elements in the sequence. These attention weights represent the importance of each input element in influencing the current output. Subsequently, the model combines the input elements with their corresponding attention weights to compute a weighted sum, effectively producing a context vector (Equation 8) that encapsulates the most relevant information from the input sequence for the current output step.

This context vector is then concatenated with the current hidden state of the LSTM layer and passed through the activation function to produce the output.

$$\alpha_{t,i} = \frac{exp\ (e_{t,i})}{\sum_{j=1}^{T} exp\ (e_{t,j})} \quad \textbf{(7)}$$

$$c_t = \sum_{i=1}^{T} \alpha_{t,i} h_i \quad \textbf{(8)}$$

Where:

- $e_{t,i}$ is a compatibility score between the decoder hidden state s_t and the encoder hidden state h_i ($e_{t,i} = score(s_t, h_i)$)
- c_t is the computed context vector.

Nevertheless, the Attention-LSTM model presents certain limitations when utilized in clinical tasks involving longitudinal data, starting from potential increased computational complexity, and training time compared to traditional LSTM models. The incorporation of attention mechanisms adds computational overhead, requiring additional resources for training and inference. Moreover, interpreting the attention weights generated by the model can be challenging, particularly in complex clinical scenarios with multiple interacting factors. Clinicians may require specialized tools or expertise to interpret and validate the attention-based predictions effectively. Additionally, the Attention-LSTM model may overfit, particularly with small or noisy datasets, necessitating careful regularization and hyperparameter tuning for improved generalization performance.

In summary, the Attention-LSTM model offers significant advantages for clinical tasks with longitudinal data, including its ability to capture temporal dependencies, adapt to variable-length sequences, and provide interpretability through attention mechanisms. However, careful consideration of its limitations, such as computational complexity and potential for overfitting, is essential for successful application in clinical practice.

Attention Bidirectional LSTM

The Attention Bidirectional Long Short-Term Memory (BiLSTM) model [47], [48], [49] is a sophisticated neural network architecture that combines the strengths of bidirectional LSTMs with an attention mechanism to process sequential data effectively. This model is particularly advantageous for tasks requiring a comprehensive understanding of input sequences, such as natural language processing, sentiment analysis, and time-series prediction. BiLSTMs are a variant of the traditional LSTM architecture that processes input sequences in both forward and backward directions. By incorporating information from past and future time steps simultaneously, bidirectional LSTMs capture contextual dependencies more comprehensively than unidirectional LSTMs.

As in the previous case, the Attention mechanism enhances the Bidirectional LSTM model's performance by enabling it to dynamically focus on relevant parts of the input sequence. This selective attention mechanism allows the model to assign different importance weights to each input element, emphasizing informative segments while deemphasizing noise or irrelevant information.

The operation of the Attention Bidirectional LSTM model involves several key steps:

1. Bidirectional LSTM Computation: The input sequence is processed simultaneously in both forward and backward directions by two separate LSTM layers. This enables the model to capture contextual information from past and future time steps for each input element.
2. Attention Computation: For each time step, the model computes attention weights by evaluating the relevance of each input element to the current context. This is typically done by comparing the current hidden state of the Bidirectional LSTM with the hidden states of all input elements using a compatibility function.
3. Context Vector Calculation: The attention weights are used to compute a weighted sum of the input elements, producing a context vector that highlights the most important information in the input sequence for the current time step.
4. Context-Enriched Representation: The context vector is concatenated with the output of the Bidirectional LSTM layer, creating a context-enriched representation of the input sequence for further processing.
5. Output Generation: The context-enriched representation is passed through additional layers, such as fully connected layers or output layers, to generate the final output of the model.

By integrating bidirectional processing with attention mechanisms, the Attention Bidirectional LSTM model can effectively capture long-range dependencies and extract relevant information from input sequences, making it a powerful tool for various clinical tasks with longitudinal data. However, this model may suffer from increased computational complexity and could be susceptible to overfitting, requiring careful regularization and hyperparameter tuning for optimal performance.

LSTM-CNN Model

The LSTM-CNN model, where an LSTM layer is connected to a CNN layer, is a hybrid neural network architecture designed to leverage the strengths of both Long Short-Term Memory (LSTM) networks and CNNs for processing sequential data, such as time-series data or sequences of text [50].

The LSTM-CNN architecture and operation comprise of:

1. Input Processing: The input data, typically represented as sequential data, is fed into the LSTM-CNN model.
2. LSTM Layer: The input sequence is initially processed by an LSTM layer. The LSTM layer consists of memory cells and gates that enable it to capture long-term dependencies and temporal patterns within the input sequence. Each LSTM cell maintains an internal

state that is updated based on the current input and the previous state, allowing it to retain relevant information over multiple time steps.

3. CNN Layer: The output of the LSTM layer is then passed to a CNN layer. The CNN layer is responsible for extracting spatial features from the sequential data representation produced by the LSTM layer. This is achieved through the application of convolutional filters across the input sequence, which detect patterns or motifs that are relevant to the task at hand.
4. Feature Extraction: As the input sequence passes through the CNN layer, it undergoes feature extraction, where local patterns or features are detected and represented in higher-level feature maps. The CNN layer typically consists of multiple convolutional and pooling layers, which enable hierarchical feature extraction and abstraction.
5. Integration of LSTM and CNN Features: The feature maps produced by the CNN layer capture spatial patterns within the input sequence, while the LSTM layer captures temporal dependencies over time. These features are then integrated or concatenated to create a combined representation that preserves both spatial and temporal information.
6. Output Generation: The combined feature representation is passed through additional layers, such as fully connected layers or output layers, for further processing and prediction. These layers may perform tasks such as classification, regression, or sequence generation, depending on the specific application of the LSTM-CNN model.

The LSTM-CNN model offers several advantages when applied to clinical tasks with longitudinal data. One advantage is its ability to capture both temporal dependencies and spatial patterns present in longitudinal data. The LSTM component excels at modeling sequential data, allowing the model to capture temporal trends and dependencies over time, which are common in clinical datasets. Meanwhile, the CNN component is effective at extracting spatial features and patterns from multidimensional data, such as medical images or time-series data with multiple features. Furthermore, the LSTM-CNN model can handle variable-length sequences of longitudinal data, making it suitable for tasks where the length of patient follow-up varies. This flexibility allows the model to adapt to different clinical scenarios without requiring fixed-length input sequences. Moreover, the LSTM-CNN model can learn hierarchical representations of data, with the CNN layers capturing low-level features and the LSTM layers capturing higher-level temporal patterns. This hierarchical representation learning enables the model to automatically extract relevant features from the input data, reducing the need for manual feature engineering.

However, the LSTM-CNN model also has certain limitations when applied to clinical tasks with longitudinal data. One limitation is the trend of overfitting, especially when dealing with small or noisy datasets. Careful regularization techniques, such as dropout and weight decay, may be necessary to prevent overfitting and improve generalization performance. Additionally, training and tuning the LSTM-CNN model can be computationally expensive, particularly when dealing with large-scale longitudinal datasets or complex model architectures. Adequate computational resources and efficient training strategies are required to mitigate this challenge. Moreover, interpreting the predictions of the LSTM-CNN model can be challenging due to its complex architecture and high-dimensional feature representations. Clinicians may require additional tools or techniques to interpret and validate the model outputs effectively.

Overall, the LSTM-CNN model offers significant advantages for clinical tasks with longitudinal data, including its ability to capture temporal and spatial patterns, handle variable-length sequences, and learn hierarchical representations. However, careful consideration of its limitations, such as overfitting and computational complexity, is essential for successful application in clinical practice.

XGBoost model

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm known for its exceptional performance across various tasks, particularly in structured data settings such as tabular data [51], [52]. Its popularity stems from its robustness, scalability, and efficiency. The key characteristics of an XGBoost model include:

- Ensemble Learning and Boosting: XGBoost belongs to the ensemble learning paradigm, specifically the boosting family. Boosting involves sequentially training a series of weak learners (models that are slightly better than random guessing) and combining their predictions to produce a strong learner (a highly accurate predictive model). XGBoost implements gradient boosting, where each new model is trained to correct the errors of the previous ones.
- Decision Trees as Base Learners: XGBoost employs decision trees as its base learners. Decision trees are hierarchical structures where each internal node represents a decision based on a feature, and each leaf node represents the predicted outcome. In XGBoost, the decision trees are typically shallow (have a limited depth) to prevent overfitting and to maintain computational efficiency.
- Gradient Boosting Framework: XGBoost builds upon the gradient boosting framework by introducing several enhancements to improve performance and speed. It minimizes a loss function by iteratively adding decision trees to the ensemble. At each iteration, the algorithm fits a new tree to the residuals (the differences between the actual and predicted values) of the previous model. The new tree is trained to predict the residuals, thereby reducing the overall error of the ensemble.
- Regularization and Control Overfitting: XGBoost incorporates regularization techniques to control overfitting and enhance generalization. It includes both L1 (Lasso) and L2 (Ridge) regularization terms in its objective function, which penalize the complexity of the model and discourage extreme parameter values. Additionally, XGBoost allows users to specify parameters such as maximum depth, minimum child weight, and subsampling rate to further control model complexity and prevent overfitting.
- Gradient Optimization: XGBoost employs an efficient and scalable algorithm for gradient optimization. It approximates the objective function using second-order Taylor expansion to speed up computation. It also utilizes a technique called "histogram-based approximation" to bin and sort feature values, reducing the memory footprint and accelerating training.
- Prediction and Scoring: Once trained, the XGBoost model can make predictions efficiently for new data instances. It aggregates the predictions of all individual trees in the ensemble and optionally applies a learning rate to control the contribution of each tree. The final

prediction is the sum of the predictions from all trees, possibly weighted by the learning rate.

XGBoost, as an efficient and scalable gradient boosting framework, offers several advantages when applied to clinical tasks with longitudinal data. One key advantage is its ability to handle complex, high-dimensional datasets commonly encountered in clinical settings. XGBoost can effectively capture non-linear relationships between input features and target variables, making it suitable for modeling the intricate dynamics present in longitudinal data. Moreover, XGBoost is well-suited for handling missing data, a common challenge in longitudinal studies, as it can automatically handle missing values during the training process without requiring imputation techniques. Additionally, XGBoost's parallel and distributed computing capabilities enable efficient training on large-scale longitudinal datasets, reducing computational time and resources. Furthermore, XGBoost provides interpretable model outputs, allowing clinicians to understand the factors driving predictions and enhancing model transparency. Its feature importance scores can help identify critical predictors of clinical outcomes, aiding in the development of personalized treatment strategies.

Despite these advantages, XGBoost also has certain limitations when applied to clinical tasks with longitudinal data. Careful regularization and hyperparameter tuning are necessary to mitigate this risk of overfitting and ensure model generalization. Additionally, XGBoost's performance may be influenced by the quality and representativeness of the longitudinal data, emphasizing the importance of data preprocessing, feature scaling and engineering to extract relevant information.

Overall, XGBoost has been proven a robust classifier in literature as it offers significant advantages for clinical tasks with longitudinal data, including scalability, interpretability, and robustness to missing values.

Unsupervised learning

K-Means

K-Means clustering is a popular unsupervised machine learning algorithm utilized for partitioning a dataset into distinct groups, or clusters, based on similarities in the data points' features [53]. K-Means clustering functions by minimizing the within-cluster variance, also known as the inertia or distortion, which quantifies the compactness of the clusters. The algorithm iteratively optimizes the cluster centroids to minimize the total distance of data points from their respective centroids. The architecture of the K-Means algorithm comprises several key components that facilitate its functionality:

1. Initialization: The algorithm begins by randomly selecting K initial cluster centroids from the dataset, where K represents the desired number of clusters.
2. Assignment Step: In this step, each data point is assigned to the nearest centroid based on a distance metric, commonly the Euclidean distance. Each data point is thereby associated with the cluster represented by the nearest centroid.

3. Update Step: After the assignment of data points to clusters, the centroids are recalculated based on the mean of all data points assigned to each cluster. These updated centroids represent the new cluster centers.
4. Convergence: The assignment and update steps are iteratively repeated until convergence is achieved. Convergence occurs when the centroids no longer change significantly between iterations or when a predefined number of iterations is reached.
5. Result: The final output of the K-Means algorithm is a set of *k* clusters, each characterized by its centroid. These clusters represent coherent groups of data points with similar features.

K-Means is widely used in various applications, including customer segmentation, image segmentation, anomaly detection, and recommendation systems. Its simplicity, efficiency, and scalability make it suitable for large datasets, although it requires the specification of the number of clusters, *k*, as a parameter. However, K-Means has limitations, such as sensitivity to initial centroid selection and its tendency to converge to local optima. Additionally, it assumes clusters of similar size and shape and is sensitive to outliers. Overall, K-Means clustering provides a versatile and efficient approach for data partitioning, enabling insights into underlying patterns and structures within datasets.

Model training

The training of Machine Learning and Deep Learning models constitutes a fundamental stage in the development of predictive systems across various domains. This process involves iteratively optimizing model parameters to minimize a chosen objective function, typically a measure of prediction error or loss. Both rely on iterative optimization algorithms, such as gradient descent variants, to update model parameters iteratively towards the direction of steepest descent. Regardless of the specific model architecture, training such models requires careful consideration of various factors, including data preprocessing, hyperparameter selection, and model architecture design.

To facilitate model development, tuning, and evaluation, the dataset is split into separate training, validation, and test sets. The training set is used for model training, the validation set for hyperparameter tuning, and the test set for final model evaluation to assess generalization performance. Splitting of training was performed per patient and per outcome, so that each set will have a similar distribution of class labels. When splitting a dataset into training and test sets, maintaining the same distribution of class labels in each set is significant for ensuring the integrity and reliability of machine learning models. This practice, known as stratified sampling, helps prevent biases in the model's performance evaluation. By preserving the proportion of different classes across the training and testing sets, the model learns to generalize better to unseen data. Additionally, maintaining the same class distribution ensures that the evaluation metrics accurately reflect the model's performance across all classes, providing more reliable insights into its effectiveness across different categories.

During training, one common problem that occurs is overfitting, meaning that the model becomes overly complex and fits the training data too closely, resulting in poor generalization

performance on unseen or test data. To handle this, techniques are applied in these experiments, like early stopping which allows the model to stop training when it begins to overfit, preventing further deterioration in performance. In addition to early stopping, regularization techniques, such as dropout, are employed, randomly deactivating a fraction of neurons during training, preventing the model from relying too much on specific features or patterns. Another effective method for handling overfitting especially in LSTM variant models is recurrent dropout, specifically applied to LSTM layers. Recurrent dropout extends the concept of dropout to the recurrent connections within LSTM units. By randomly setting a fraction of the recurrent connections to zero during training, recurrent dropout prevents LSTM units from memorizing the training data excessively, encouraging better generalization to unseen sequences. This regularization technique helps LSTM models to learn more robust representations of sequential data, reducing overfitting and improving overall performance. When combined with early stopping and other regularization methods, such as dropout between layers, recurrent dropout enhances the model's ability to generalize and make accurate predictions on sequential tasks.

Model optimization

In accordance with model training, optimizing Machine Learning and Deep Learning models is a critical aspect of developing effective predictive systems across various domains. This process involves fine-tuning model parameters, hyperparameters, and training procedures to enhance model performance and generalization ability.

During hyperparameter search, grid search is commonly used to explore the hyperparameter space and identify optimal configurations that maximize model performance. Grid search is an exhaustive search technique, meaning it evaluates all possible combinations of hyperparameters within the specified grid. While grid search is simple and easy to implement, it can be computationally expensive, especially when dealing with a large number of hyperparameters or when the hyperparameter space is vast. In this thesis, custom grid search functions are implemented to ensure consistency of sub-datasets distribution and specific evaluation schema of models. For Deep Learning models implemented in this thesis, hyperparameters include:

- **Learning Rate**: The learning rate determines the size of the step taken during optimization and significantly influences model convergence and stability.
- **Optimizers**: Optimizers play a vital role in updating model parameters during training. Specifically, three optimizers will be examined in these experiments, namely:
    - Adam (Adaptive Moment Estimation): Adam is a popular optimization algorithm maintaining adaptive learning rates for each parameter by computing exponentially decaying averages of past gradients and squared gradients. Adam adapts the learning rate for each parameter individually, making it well-suited for a wide range of machine learning tasks.
    - Stochastic Gradient Descent (SGD): SGD is a classic optimization algorithm that updates model parameters using the gradient of the loss function computed on a single training example or a mini batch of examples.

- RMSProp (Root Mean Square Propagation): RMSProp is an adaptive learning rate optimization algorithm that addresses the diminishing learning rate problem by using a moving average of squared gradients. RMSProp divides the learning rate by the exponentially decaying average of squared gradients, effectively scaling down the learning rate for frequently occurring features and scaling up the learning rate for infrequently occurring features.

Hyperparameter search involves tuning optimizer-specific parameters, such as momentum, learning rate decay, and epsilon, to optimize model training.

- **LSTM Units**: The number of LSTM units in a layer directly influences the model's memory capacity and representational power. A higher number of LSTM units enable the model to capture more complex temporal patterns and dependencies, thereby increasing its ability to learn from sequential data. However, this comes at the cost of increased computational complexity and potential overfitting, especially when dealing with limited training data. On the other hand, a lower number of LSTM units may lead to underfitting, where the model lacks the capacity to capture intricate temporal relationships within the data. Therefore, finding the optimal number of LSTM units involves a trade-off between model complexity and generalization performance.
- **CNN filters**: The number of filters in a Convolutional Neural Network layer determines the depth or the number of features the network can learn from the input data. Each filter acts as a feature detector, scanning across the input image or feature map to detect specific patterns or features.
- **Dense Units**: Dense units refer to the number of neurons or units in each fully connected layer of a neural network. The number of dense units in a neural network's hidden layers determines the model's capacity to learn complex patterns and representations from the input data. Increasing the number of dense units in a hidden layer allows the model to learn a richer set of features, potentially improving its ability to represent complex relationships within the data. However, this also increases the model's parameter count and computational complexity, making it more prone to overfitting, especially when dealing with small datasets. Conversely, reducing the number of dense units can help mitigate overfitting and improve generalization performance, especially in scenarios where the training data is limited.
- **Dropout Rates**: Determining the proportion of neurons to deactivate at each training iteration, to optimize model generalization performance.
- **Number of Epochs**: Defining the number of times the entire training dataset is passed forward and backward through the model during training.
- **Batch Size:** The batch size specifies the number of samples processed before updating the model parameters.

Model validation

Performance Metrics

Model evaluation metrics play a pivotal role in assessing the efficacy and robustness of AI models in various domains, including healthcare applications such as predicting critical events in Multiple Myeloma. In supervised learning, several key evaluation metrics commonly used to gauge the performance of predictive models include accuracy, area under the receiver operating characteristic curve (AUC), precision, recall, F1 score, and area under the precision-recall curve (AUPRC). Accuracy (Equation 9) is a fundamental metric that measures the overall correctness of predictions made by a model. It quantifies the proportion of correctly classified instances among the total number of instances in the dataset. While accuracy provides a general indication of model performance, it may not be suitable for imbalanced datasets, where the majority class dominates the predictions, only in case the imbalance is handled [54]. Area under the receiver operating characteristic curve is a widely used metric for evaluating the discriminative power of binary classifiers. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold values. AUC represents the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance by the classifier. A higher AUC value indicates better discriminative performance of the model. Precision (Equation 10) quantifies the proportion of true positive predictions among all instances predicted as positive by the model. It focuses on the accuracy of positive predictions and is particularly relevant in scenarios where false positives have significant consequences. Precision is computed as the ratio of true positives to the sum of true positives and false positives. Recall (Equation 11), also known as sensitivity or TPR, measures the proportion of true positive instances that are correctly identified by the model. It gauges the model's ability to capture all positive instances in the dataset, regardless of false negatives, being the most valuable metric in predicting critical events as in this thesis. Recall is calculated as the ratio of true positives to the sum of true positives and false negatives. A high sensitivity indicates that the model is good at detecting cases of clinical interest, minimizing the number of false negatives (cases where critical clinical events are missed) [55]. Sensitivity is an important metric, especially in medical applications, where missing positive cases (relapse, mortality and treatment response in this case) can have serious consequences. F1 score (Equation 12) is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It takes into account both false positives and false negatives and is especially useful in situations where the class distribution is imbalanced. F1 score ranges from 0 to 1, with higher values indicating better model performance. AUPRC quantifies the trade-off between precision and recall across different classification thresholds. Unlike AUC, which focuses on the TPR and FPR, AUPRC considers precision and recall directly. A higher AUPRC value reflects better model performance, especially in scenarios where class imbalance is prevalent.

In summary, performance evaluation metrics such as accuracy, AUC, precision, recall, F1 score, and AUPRC provide valuable insights into the effectiveness of machine learning models in predicting critical events in Multiple Myeloma. By comprehensively assessing different aspects of model performance, these metrics aid in model selection, optimization, and interpretation, ultimately facilitating the development of more accurate and reliable predictive models for clinical decision support.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(9)}$$

$$Precision = \frac{TP}{(TP + FP)} \quad \textbf{(10)}$$

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(11)}$$

$$F1\ score = \frac{2\ x\ (Precision\ x\ Recall)}{(Precision + Recall)} \quad \textbf{(12)}$$

TP: Number of True Positives
FP: Number of False Positives
TN: Number of True Negatives
FN: Number of False Negatives

For each prediction task these numbers are further analyzed into:

Table 2. Description of results with positive and negative classifications

Prediction task	**TP**	**FP**	**TN**	**FN**
Treatment response	Cases where the model correctly predicts that a patient responds to the treatment (positive class) when they actually do respond according to ground truth labels.	Cases where the model incorrectly predicts that a patient responds to the treatment (positive class) when they actually do not respond.	Cases where the model correctly predicts that a patient does not respond to the treatment (negative class) when they actually do not respond.	Cases where the model incorrectly predicts that a patient does not respond to the treatment (negative class) when they actually do respond.
Relapse	Cases where the model correctly predicts that a patient will experience a relapse.	Cases where the model incorrectly predicts that a patient will experience a relapse when, in reality, they do not.	Cases where the model correctly predicts that a patient will not experience a relapse.	Cases where the model incorrectly predicts that a patient will not experience a relapse when, in reality, they do.
Mortality	Cases where the model correctly predicts that a patient has died (positive class) and the patient has indeed died according to the ground truth.	Cases where the model incorrectly predicts that a patient has died (positive class) when the patient has not died.	Cases where the model correctly predicts that a patient has not died (negative class) and the patient has indeed not died according to the ground truth.	Cases where the model incorrectly predicts that a patient has not died (negative class) when the patient has indeed died.

In the case of clustering, and specifically when using K-Means, determining the optimal number of clusters (*k*) is crucial for meaningful analysis. Two common methods for evaluating the appropriateness of the number of clusters applied in these experiments are the Silhouette Analysis and the Elbow Method. The Silhouette Analysis which calculates the silhouette score, quantifies how well each data point fits its assigned cluster relative to other clusters, with values ranging from -1 to 1. A higher silhouette score indicates better cluster separation and cohesion. By calculating the average silhouette score across all samples, the optimal *k* value can be identified as the one that maximizes this metric, indicating the presence of well-defined clusters.

Conversely, the Elbow Method involves plotting the sum of squared distances (inertia) from each point to its cluster center against different values of k. As *k* increases, inertia decreases, indicating improved clustering. The objective is to identify the "elbow point" on the plot, where the rate of decrease sharply slows down. This point signifies the optimal number of clusters, striking a balance between cluster compactness and model complexity. By integrating these approaches, analysts can systematically determine the most suitable *k* value, enhancing the interpretability and validity of the K-Means clustering outcomes.

Cross-Validation

Using a patient-wise cross-validation schema for ML model validation is a robust approach, particularly in healthcare and clinical research settings. Unlike traditional cross-validation methods that randomly partition data into folds [56], [57], patient-wise cross-validation preserves the temporal and individual dependencies present in longitudinal patient data. In this schema, patients are grouped into folds such that all data points from a single patient are included in the same fold. This ensures that the model is trained and evaluated on patient-level data, mimicking real-world scenarios where the model must generalize to new patients rather than individual data points. By maintaining patient-wise separation, this approach provides a more realistic estimate of model performance and helps prevent data leakage or overfitting that can occur with other cross-validation methods. Additionally, patient-wise cross-validation facilitates the assessment of model robustness across diverse patient populations and enables more meaningful comparisons between different algorithms or feature sets. Overall, adopting a patient-wise cross-validation schema enhances the reliability and interpretability of forecasting models in healthcare applications.

Explainability

Explainability in Machine Learning refers to the capacity of a model to provide interpretable and transparent insights into its decision-making process [58], [59]. Therefore, explainability is paramount in healthcare settings, where decisions directly impact patient outcomes. Clinicians and stakeholders require transparent models that provide actionable insights and justify recommendations. In the context of MM prediction, explainability enables clinicians to understand the rationale behind a model's predictions, identify influential features, and validate the clinical relevance of predictive biomarkers [4]. By demystifying complex ML algorithms, explainable

models foster collaboration between clinicians and data scientists, facilitating knowledge exchange and promoting evidence-based decision-making.

Several explainability techniques have been developed to elucidate the inner workings of ML models. Feature importance analysis, such as permutation importance and SHAP (SHapley Additive exPlanations), quantifies the contribution of each input feature to the model's predictions [60], [61]. By ranking features based on their impact on model performance, clinicians gain insights into the biological relevance of predictive biomarkers and their associations with disease progression in MM. Specifically, the core idea behind SHAP values is to quantify the impact of each feature on the model's predictions by evaluating the change in predictions when considering the feature's absence or presence, while accounting for all possible combinations of features. This is achieved by leveraging SHapley values from cooperative game theory, which attribute a value to each player (feature) based on their marginal contribution to the overall payoff (prediction). In practice, computing SHAP values involves generating a dataset of perturbed instances by systematically removing or altering feature values, and then evaluating the model's predictions on these perturbed instances. The difference between the model's predictions on the original and perturbed instances reflects the contribution of each feature to the prediction.

Experimental Design

In this section, the implementation procedures of all prediction tasks are further analyzed, to provide a comprehensive explanation of the methodology followed to set up the prediction task, as well as train and validate the machine learning models.

Relapse prediction

Methodology

Setting up the prediction task of early relapse in two prediction horizons, namely at ‘Year 1’ and at ‘Year 5’, corresponding time intervals were extracted in each case. The distribution of relapse across all time intervals in the dataset is depicted in Figure 8, showing the highest percentages of relapse between ‘Year 1’ and ‘Year 2’.

Specifically, in the case of predicting early relapse at ‘Year 1’, in order to keep all longitudinal information available in the dataset, visit intervals of ‘Baseline’, ‘Month 3’, ‘Month 6’ and ‘Month 9’ were kept as the total observation window, whose data were utilized as input. Similarly, in the case of predicting early relapse at ‘Year 5’, visit intervals up to ‘Year 1’ were included in input data (Figure 9). This methodology aimed at implementing both ‘short-term’ predictions, considering the 3-month interval for predicting relapse at the first year, while also implementing a more ‘long-term’ prediction by utilizing first year’s data to predict relapse at ‘Year 5’.

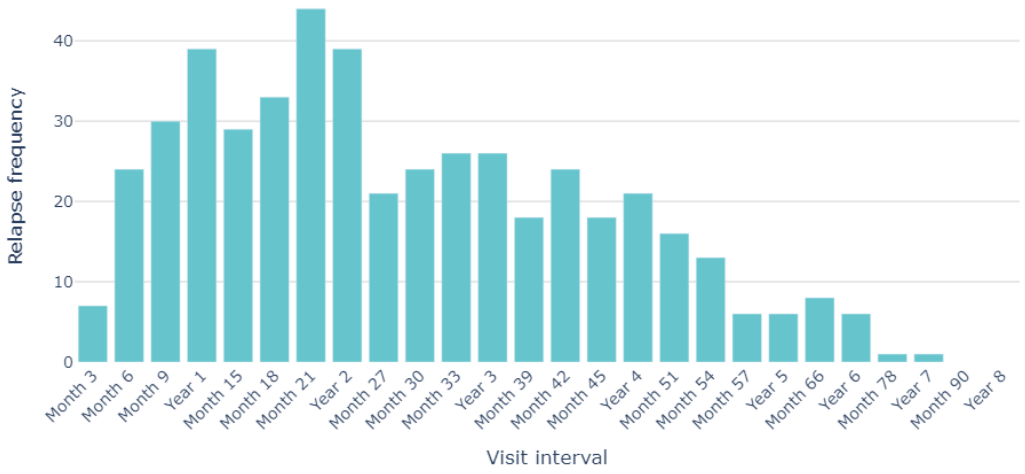

Figure 8. Distribution of relapse recordings across visit intervals.

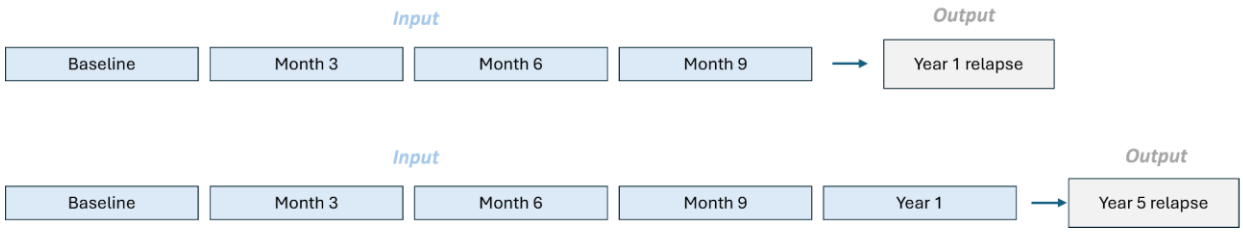

Figure 9. Input & Output intervals for relapse prediction.

This prediction task is defined as a binary classification task, where the positive class (1) represents the event of relapse, while the negative class (0) represents the absence of relapse event. The DL model utilized in predicting relapse is the Attention LSTM model, due to its capability to selectively focus on relevant temporal information, effectively capturing long-term dependencies and patterns within sequential data. For optimization of model hyperparameters, a custom grid function was developed, in order to identify the optimal combination of these hyperparameters leading to the best performance metrics results. The grid search hyperparameter space is analyzed in Table 3.

Furthermore, the experiments for this prediction task included running the model multiple times (i.e. 10 times), since variations in results are observed due to various reasons. Specifically, neural networks, including LSTMs, are initialized with random weights, leading to differences in model performance across runs, which helps assess stability and understand outcome ranges. Additionally, LSTMs are sensitive to initialization and hyperparameters, so multiple runs allow for robustness evaluation and optimal configuration identification. Real-world datasets are inherently variable, and multiple runs assess model response to this variability. This model is trained on the original input dataset including all features, as well as on a sub-dataset extracted with 50 features, from applying feature selection with Random Forest.

Treatment Response prediction

Methodology

This task of predicting treatment response is implemented using three different modeling approaches. The outcome variables of interest included ‘treatment response’, reflecting the treatment response label at each visit. Treatment response in the MMRF dataset is originally represented with 6 different labels, from most responsive to least responsive, namely 'Stringent Complete Response', 'Complete Response', 'Very Good Partial Response', 'Partial Response', 'Stable Disease', and 'Progressive Disease'. The original distribution of treatment response is depicted in Figure 2. After initial experimentation across all approaches, to achieve well-performing results, the six-class variables are converted initially into a three-class variable, with labels ‘No Response’ (‘Stable Disease’ & ‘Progressive Disease’), ‘Partial Response’ (‘Very Good Partial Response’ & ‘Partial Response’) and ‘Complete Response’ (‘Stringent Complete Response’ & ‘Complete Response’). For two out of three methodologies implemented, this task was converted into a binary classification task, converting the three-class treatment response variable into a binary variable, namely ‘Not Responsive’ (‘No Response’) and ‘Responsive’ (‘Partial Response’ & ‘Complete Response’).

Regarding the same methodologies, since first line treatment is clinically a significant factor of the disease trajectory indicating early relapse and decrease of overall survival, the treatment response prediction at both prediction horizons is focused on first-line treatment predictions. The distribution of response for first-line treatments is analyzed in Figure 10.

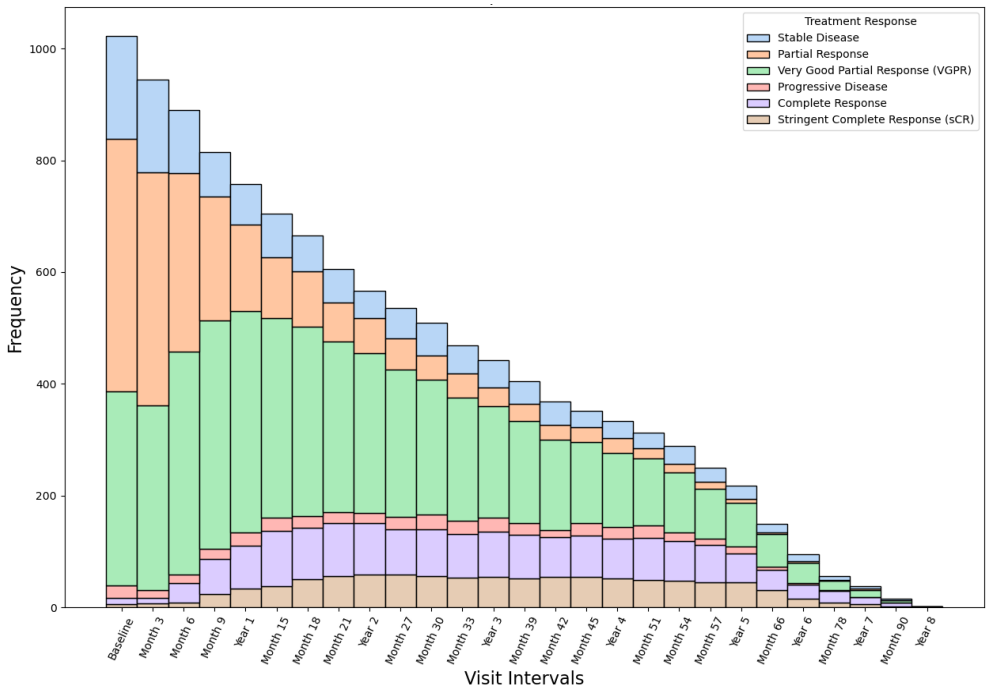

Figure 10. First-line treatment response distribution across visit intervals.

Methodology 1(TR-M.1). Treatment response prediction at Year 1 and Year 5 utilizing baseline data

This methodology focuses on predicting first-line treatment responses of patients at 'Year 1' and 'Year 5' since recruitment, using as input data only the information available at the first visit (Baseline)(Figure 11). The outcome variable is formulated as a binary variable, where the positive class represents the being 'Responsive' to first-line treatment.

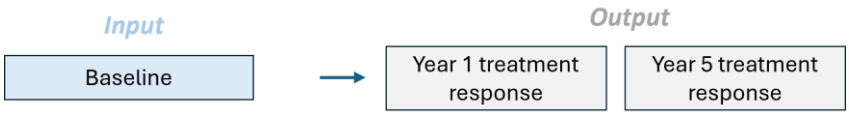

Figure 11. Input & Output intervals for treatment response (TR-M.1)

For this implementation, XGBoost model was employed, to effectively identify significant features and interactions that differentiate between responders and non-responders. By tuning hyperparameters such as learning rate, max depth, and number of trees, the model can be optimized to achieve high accuracy and predictive power (Table 3). Additionally, its built-in cross-validation capabilities and feature importance metrics provide valuable insights into the model's performance and the most influential factors in predicting treatment outcomes.

Methodology 2(TR-M.2). Treatment response prediction at Year 1 and Year 5 (Longitudinal approach)

Similarly with TR-M.1 approach, in this case data from Baseline until 'Month 9' which is the last visit interval before 'Year 1', are used for training the AI models (Figure 12). Thereby, patient trajectory is captured in an effort to extract significant patterns reflecting treatment response. The outcome is handled again as a binary variable representing responsiveness for first-line treatment, using advanced neural networks, namely Attention LSTM and LSTM-CNN models. As referenced in AI Methodologies section, these models leverage the sequential learning capabilities of LSTM, and were further optimized to maximize these capabilities (Table 3).

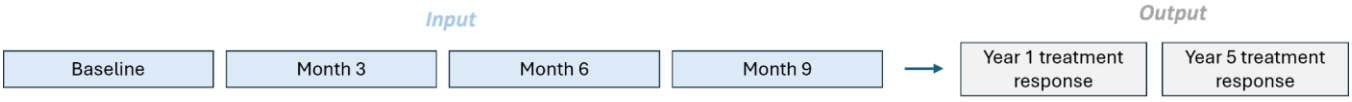

Figure 12. Input & Output intervals for treatment response (TR-M.2)

Methodology 3(TR-M.3). Treatment response prediction every 6 months ahead (Sliding windows Approach)

In this approach, the sliding window method is implemented, as it occurs as a popular technique used in time series analysis and forecasting. This approach involves creating a moving window of a fixed size, in this case, six months, that slides over the time series data. Each window captures a sequence of data points corresponding to six months of patient history, including variables like medical test results, medication adherence, and other relevant clinical metrics. The LSTM-variant models, which are adept at learning from sequential data due to the LSTM model's ability to retain long-term dependencies, are trained on these windows to learn patterns and temporal relationships in the data. Specifically, the Attention-LSTM, Attention BiLSTM and LSTM-CNN models were examined for their effectiveness in handling this binary classification problem, by choosing optimal hyperparameters (Table 3). By sliding the window one step at a time, specifically every 6 months in this case, the model generates predictions for the treatment response at each step. This method allows the model to continuously update its understanding and predictions as new data becomes available, providing a dynamic and up-to-date forecasting tool (Figure 13).

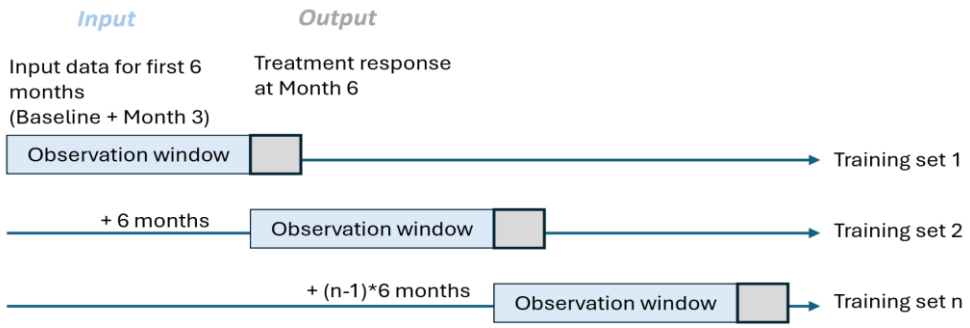

Figure 13. Input & Output intervals for treatment response (TR-M.3)

Mortality Risk prediction

Methodology

For this prediction task, a prediction window of five years was selected since this time interval is considered indicative and is commonly utilized for defining survival rates.

Methodology 1(MR-M.1). Mortality risk prediction within 5-years since baseline

Regarding this approach, the data used for training the model remained the information collected at baseline, in an effort to reflect real-world scenarios where given some initial patient profile, mortality risk could be calculated (Figure 14). This task is defined as a binary classification problem, where the positive class ('1') stands for mortality within 5-years.

By leveraging high-dimensional baseline clinical data, ML models such as XGBoost can offer significant predictive power and interpretability. XGBoost is employed due to its robustness, scalability, and ability to handle missing values. The algorithm's hyperparameters, such as the learning rate, maximum depth of trees, subsample ratio, and regularization parameters, are optimized using a 5-fold cross-validation technique to prevent overfitting and enhance

generalization (Table 3). The final optimized XGBoost model is validated on an independent test dataset to assess its generalizability and robustness. As a final step, a risk score was extracted based on the probabilities of predicting the negative class, defined as the mortality risk.

Figure 14. Input & Output intervals for mortality risk prediction (MR-M.1)

Methodology 2(MR-M.2). Mortality risk prediction within 5-years for patient clusters

Comparably, this prediction task was implemented utilizing baseline data of each patient cluster derived from K-Means clustering application. Specifically, utilizing the K-Means algorithm for identifying patient clusters based on RNA-sequencing data involves partitioning patients into distinct groups with similar gene expression profiles. This approach helps uncover underlying biological patterns and patient subtypes, leading to more patient-tailored clinical outcomes (Figure 15).

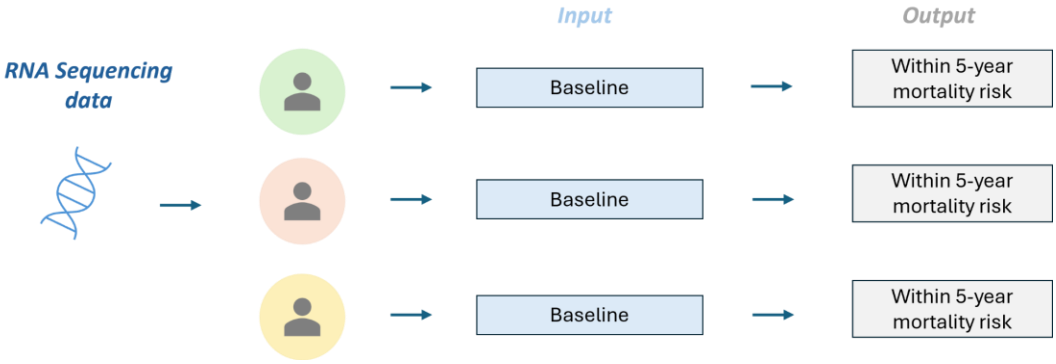

Figure 15. Input & Output intervals for mortality risk prediction (MR-M.2)

Before applying K-Means clustering algorithm, Principal Component Analysis was utilized, firstly to reduce the dimensionality of the data, transforming high-dimensional gene expression profiles into a lower-dimensional space while retaining most of the variability in the data. Thus, easier visualization of the clustering results is allowed, to visually assess the separation and cohesion of the identified patient clusters. Moreover, PCA can reveal the most significant features that contribute to the variance in the data, enhancing the robustness of the clustering process. The total number of PCA components was determined utilizing the Variance Explained and Scree plot. Variance Explained quantifies the proportion of total variance each principal component accounts for, offering a numerical criterion for component selection. In contrast, the Scree Plot visually

represents the variance explained by each component, aiding in identifying the optimal cutoff point.

To compare differences between the clusters derived in a population, statistical tests such as the t-test and Chi-square test are employed [62]. The t-test, and specifically the independent t-test, is particularly useful for comparing the means of continuous variables between the two clusters, assessing if there is a statistically significant difference in means. This test, suited for normally distributed data and small sample sizes. Conversely, the Chi-square test examines the association between categorical variables within the clusters, evaluating if observed frequencies in a contingency table significantly differ from expected frequencies. Ideal for categorical data analysis, this test includes assessments of independence and goodness of fit, offering insights into relationships without the assumption of normality. The aim of employing these tests is to gain a comprehensive understanding of the unique characteristics and relationships within the clusters, facilitating informed decision-making and deeper insights into the population dynamics.

After the final patient clusters are determined, *Methodology 1(MR-M.1). Mortality risk prediction within 5-years since baseline* is applied separately on each cluster, to evaluate mortality risk prediction capability of tailored ML models.

For all experiments, model optimization was performed, requiring an extensive search of a hyperparameter space. All values of hyperparameters examined are given below in Table 3.

Table 3. Hyperparameter space of all models.

AI/ML model	Hyperparameter space
Attention LSTM & Attention BiLSTM	Batch size: [4,8,16,32] Learning rate: [1e-2,1e-3, 1e-4, 1e-5] Number of epochs: [40,50,100,200] Optimizer: [RMSProp, Adam, Stochastic Gradient Descent] Units of LSTM layer: [16,32,40,64] Units of Dense layer: [16,32,64,128,256] Recurrent dropout of LSTM layer: [0,0.4] Dropout rate: [0, 0.1, 0.3,0.5]
LSTM-CNN	Batch size: [4,8,16,32] Learning rate: [1e-2,1e-3, 1e-4, 1e-5] Number of epochs: [40,50,100,200] Optimizer: [RMSProp, Adam, Stochastic Gradient Descent] Units of LSTM layer: [16,32,40,64] Units of CNN layer: [16,32] Units of Dense layer: [16,32,64,128,256] Recurrent dropout of LSTM layer: [0,0.4] Dropout rate: [0, 0.1, 0.3,0.5]
XGBoost	Number of estimators: [50,100,300,500] Learning_rate: [0.01, 0.001, 0.1]

	Maximum depth: [3,5,7] alpha parameter: [0, 0.1, 1] Minimum child weight: [1, 3, 10] lambda parameter: [1, 2] Subsampling factor: [0.6, 1.0] Column Subsampling by tree: [0.6, 1.0]

Results

The results of these thesis experiments reveal significant insights into the performance and effectiveness of the proposed models across various experimental setups and prediction tasks.

Relapse prediction

This task results in overall good models' performance, especially at predicting relapse in the first year since study enrollment and treatment, indicating its superior performance in capturing and leveraging temporal dependencies within the data in predicting patient relapse at both prediction horizons. Evaluation metrics results over the 10 training iterations and evaluation on the test dataset are given in Table 4. Each model is trained once on the entire input dataset, as well as on the predictors selected with Random Forest Classifier. These predictors include percentage of plasma cells in bones, flow cytometer measurements in bone marrow, LDH (ukat/L), Serum IgA, Serum IgG, Beta 2 Microglobulin (mcg/mL), Serum Lambda (mg/dL), Serum M-component, 24 hr Urine Total Protein (g/24 hr), Creatinine (umol/L), Proliferation index, answers from QoL questionnaires such as treatment side effects, maximum lines of therapy, best treatment response, treatment cycle, stem cell autologous transplant flag, age, weight, status of MM, patient death.

The final model's architecture is depicted below (Figure 16). Training and validation procedures for both accuracy and loss are plotted in Figure 17-Figure 18, along with the corresponding ROC curves at each iteration (Figure 19-Figure 20).

Overall, due to the imbalance, and the downsizing of data especially after undersampling the original number of patients, overfitting is evident during model training. Despite this trend, the model in the case of feature selection prior training, managed to perform well on the prediction task.

Table 4. Results of relapse prediction

AI/ML Model	Prediction window	Final hyperparameters	Imbalance ratio	Performance metrics (Mean, std (%))
Attention LSTM	Year 1	Batch size: 16 Learning rate: 0.0001 Optimizer: Adam Number of epochs: 100 Dropout rate: 0.2 LSTM units: 32	Negative samples: 300 Positive samples: 98	**Accuracy**: 63.6, 5.2 **Recall:** 49.3, 6.0 **Precision:** 48.6, 8.3 **F1-score:** 48.6, 6.9

		Dense units: 32		
Attention LSTM (with feature selection)	Year 1	Batch size: 16 Learning rate: 0.0001 Optimizer: Adam Number of epochs: 100 Dropout rate: 0.3 LSTM units: 32 Dense units: 32	Negative samples: 300 Positive samples: 98	**Accuracy**: 81.6, 5.4 **Recall:** 69.6, 8.9 **Precision:** 80.4, 9.6 **F1-score:** 71.4, 10.0
Attention LSTM	Year 5	Batch size: 8 Learning rate: 0.0001 Optimizer: Adam Number of epochs: 100 Dropout rate: 0.2 LSTM units: 32 Dense units: 32	Negative samples: 100 Positive samples: 41	**Accuracy**: 65.9, 9.6 **Recall:** 57.0, 10.7 **Precision:** 59.9, 16.45 **F1-score:** 56.6, 12.3
Attention LSTM (with feature selection)	Year 5	Batch size: 8 Learning rate: 0.0001 Optimizer: Adam Number of epochs: 80 Dropout rate: 0.3 LSTM units: 32 Dense units: 32	Negative samples: 100 Positive samples: 41	**Accuracy**: 68.0, 8.3 **Recall:** 58.0, 9.2 **Precision:** 62.2, 17.4 **F1-score:** 56.8, 11.0

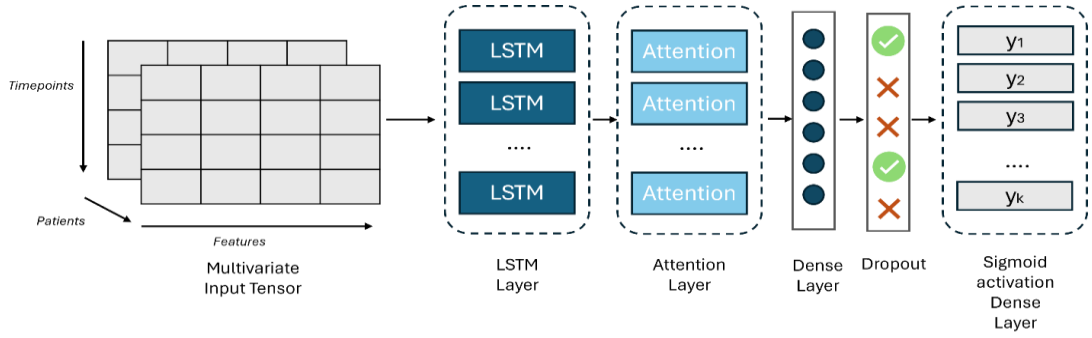

Figure 16. Attention-LSTM architecture for relapse prediction.

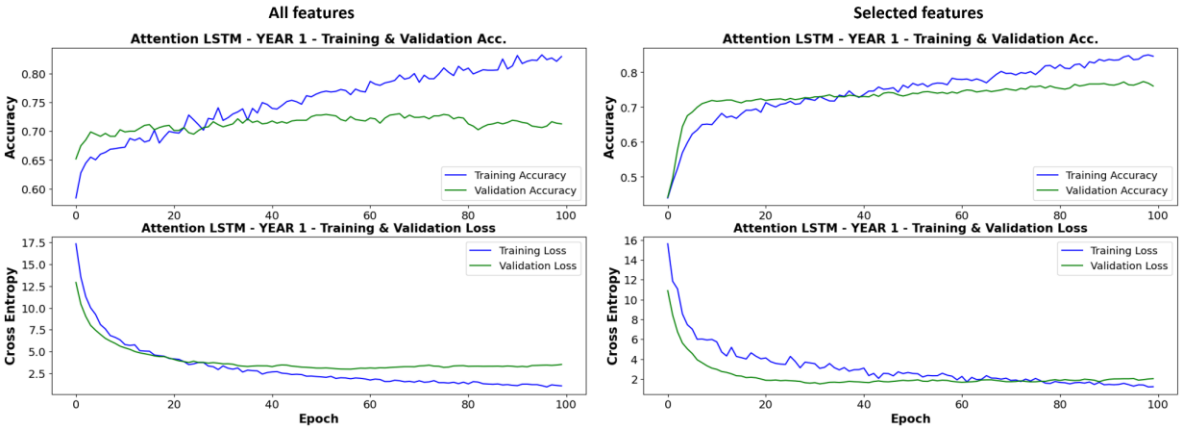

Figure 17. (Left-top) Training and validation accuracy curves with all features, (Left-down) Training and validation loss curves for all features, (Right-top) Training and validation accuracy curves with selected features, (Right-down) Training and validation loss curves for selected features. Relapse prediction horizon is set at one year.

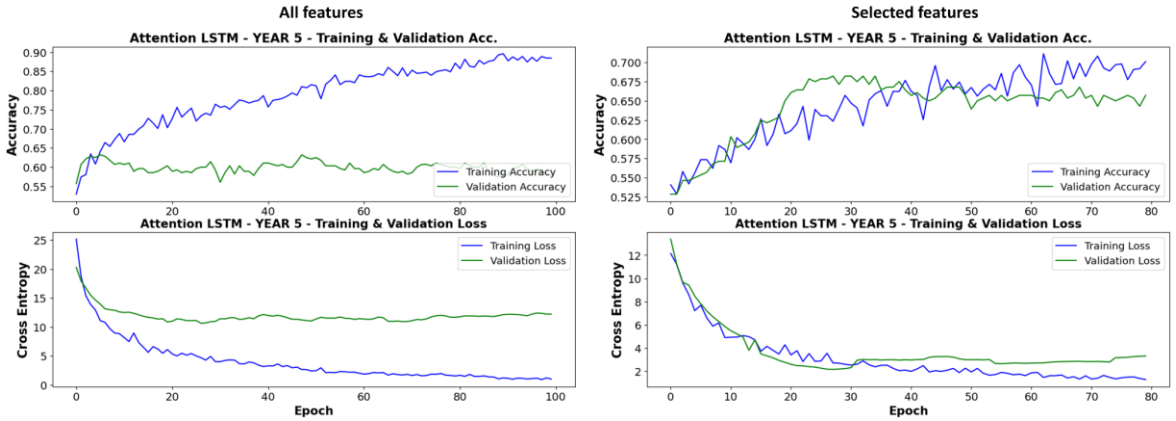

Figure 18. (Left-top) Training and validation accuracy curves with all features, (Left-down) Training and validation loss curves for all features, (Right-top) Training and validation accuracy curves with selected features, (Right-down) Training and validation loss curves for selected features. Relapse prediction horizon is set at five years.

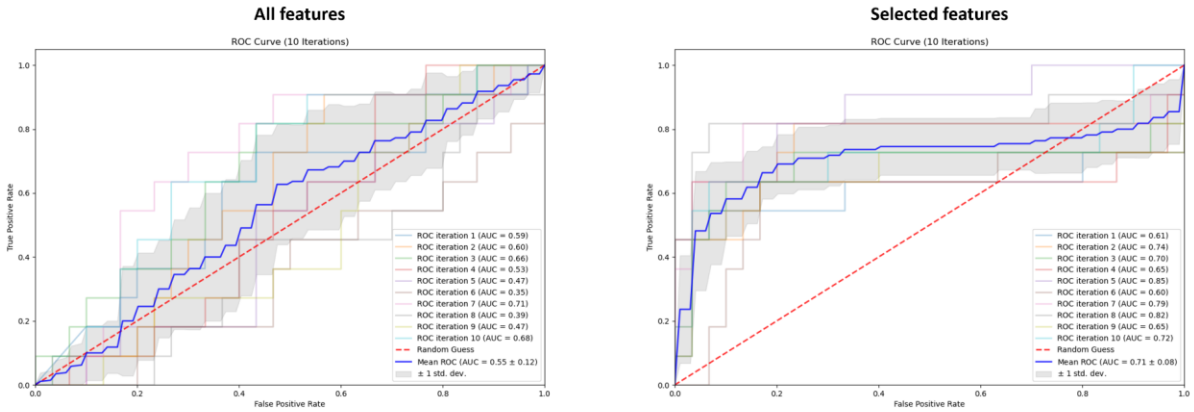

Figure 19. ROC Curves for 10 training iterations for prediction horizon set at one year with all features (Left) and selected features (Right)

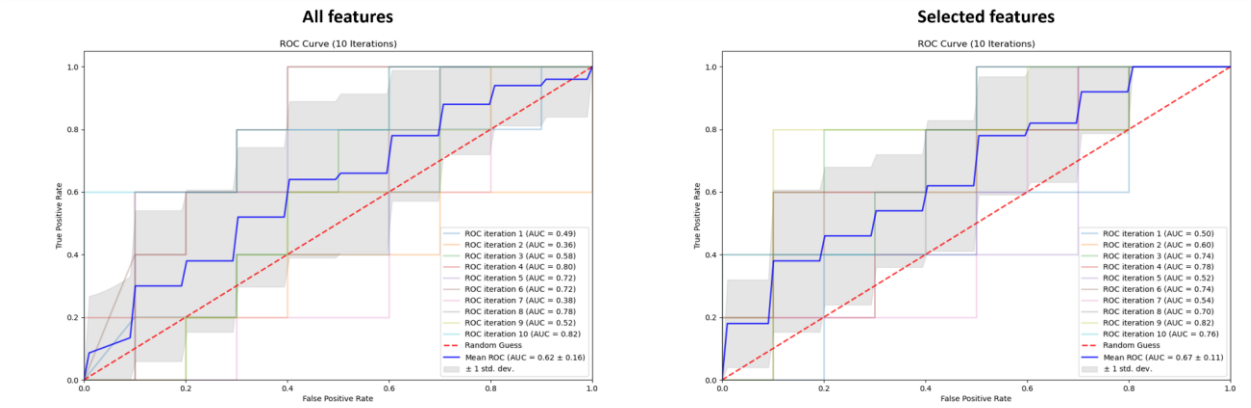

Figure 20. ROC Curves for 10 training iterations for prediction horizon set at five years with all features (Left) and selected features (Right)

Treatment Response prediction

TR-M.1. Treatment response prediction at Year 1 and Year 5 utilizing baseline data

The models' performance of this prediction task with baseline data yields excellent performance results (Table 5). The XGBoost model demonstrates strong performance across several key metrics, indicating its effectiveness in the given task. For predicting treatment response at 'Year 1', with an accuracy of 91.6%, the model correctly predicts the outcomes for the majority of instances. The recall of 88.4% suggests that the model successfully identifies a high proportion of true positive cases, while correctly also identifying true positives with precision being at 89.4%.. These results indicate that the XGBoost model is reliable and well-suited for the classification task at hand. Similarly, for predicting first-line treatment response at 'Year 5', the model's prediction ability across all evaluation metrics is very strong. ROC Curve and Precision-Recall curves are given below for each prediction horizon (Figure 21-Figure 24).

According to the feature importance and SHAP values plot for predicting treatment response at 'Year 1' (Figure 22-Figure 23), the most influential features relate to information regarding response, such as treatment response cycle, response duration, best treatment response cycle, which is expected behavior since the outcome of interest is to predict the response label. Regarding 'Year 5' prediction horizon, similar importance is extracted (Figure 25-Figure 26), however more lab measurements values are presented with importance. Some important lab measurements extracted are C-Reactive Protein (mg/dL), plasma cells in Bones (%), Flow cytometer aneuploidy population (%), Hemoglobin (mmol/L), Serum M Protein (g/dL) and 24-hour Urine M Protein (g/24 hr).

Table 5. TR-M.1 prediction results

AI/ML Model	Prediction window	Final hyperparameters	Imbalance ratio	Performance metrics (%)
XGBoost	Year 1	Number of estimators: 50 Learning rate: 0.001 Subsampling factor: 1.0 Column sampling factor: 0.6	Negative samples: 91	**Accuracy**: 91.6 **Recall:** 88.4 **Precision:** 89.4

		Maximum depth: 7 Minimum child weight: 1.0 lambda parameter: 1.0 alpha parameter: 0	Positive samples: 500	**F1-score:** 88.9
XGBoost	Year 5	Number of estimators: 50 Learning rate: 0.001 Subsampling factor: 1.0 Column sampling factor: 0.6 Maximum depth: 7 Minimum child weight: 1.0 lambda parameter: 1.0 alpha parameter: 0	Negative samples: 36 Positive samples: 150	**Accuracy**: 87.5 **Recall:** 78.9 **Precision:**88.5 **F1-score:** 81.9

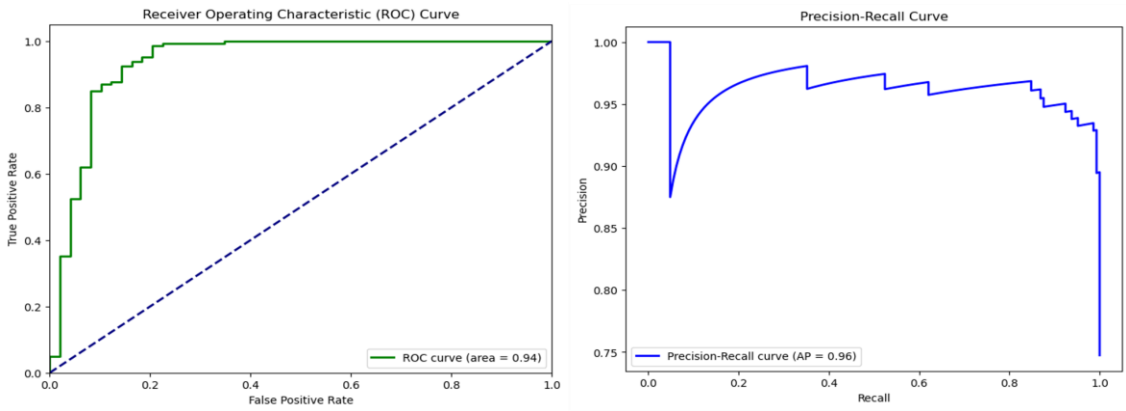

Figure 21. (Left) ROC curve and (Right) Precision-Recall curve for TR-M.1 with prediction horizon set at year 1.

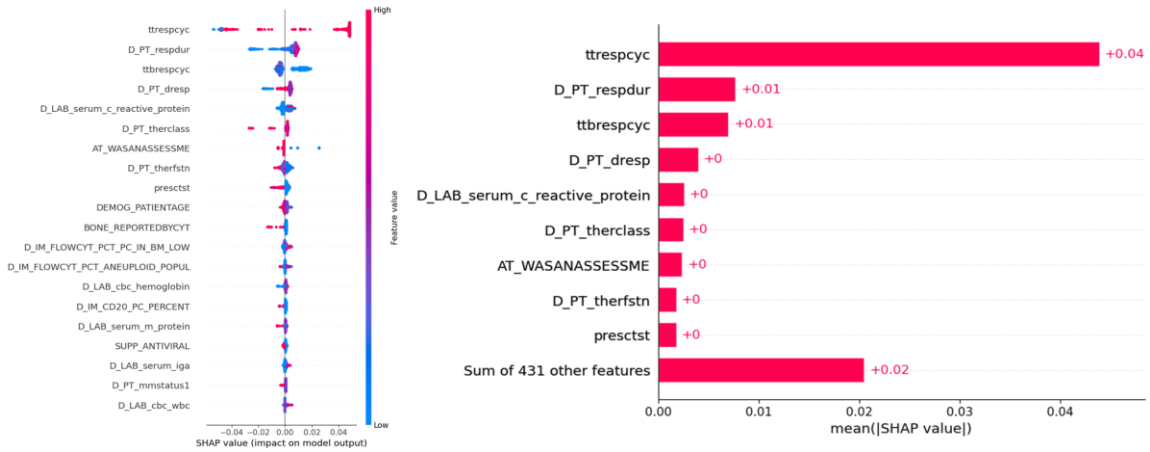

Figure 22. SHAP explainability for TR-M.1 with prediction horizon set at year 1.

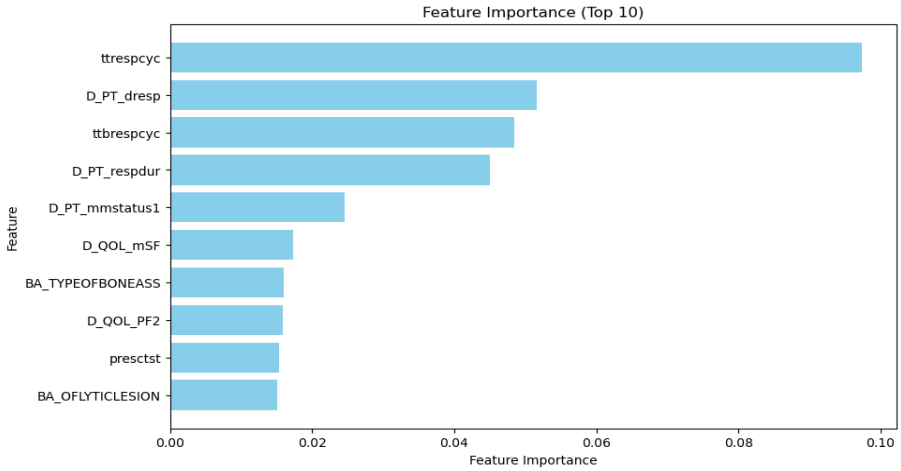

Figure 23. Feature importance of XGBoost model for TR-M.1 with prediction horizon at year 1.

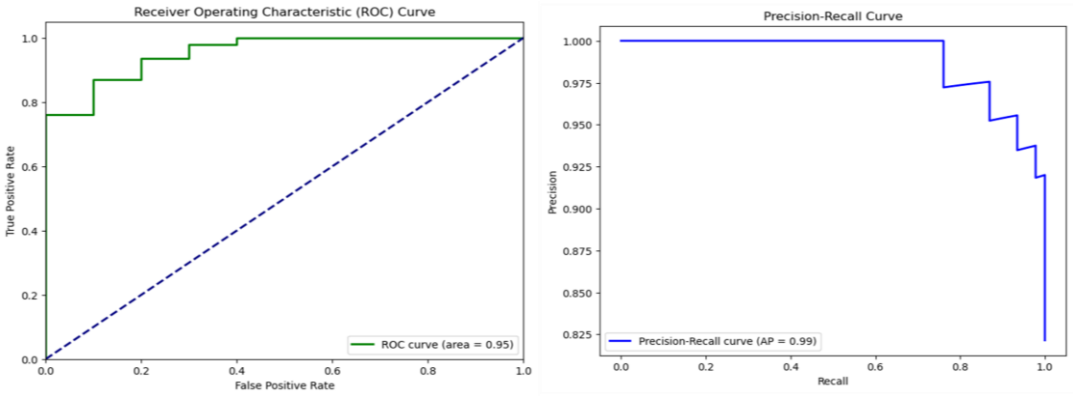

Figure 24. (Left) ROC curve and (Right) Precision-Recall curve for TR-M.1 with prediction horizon set at year 5.

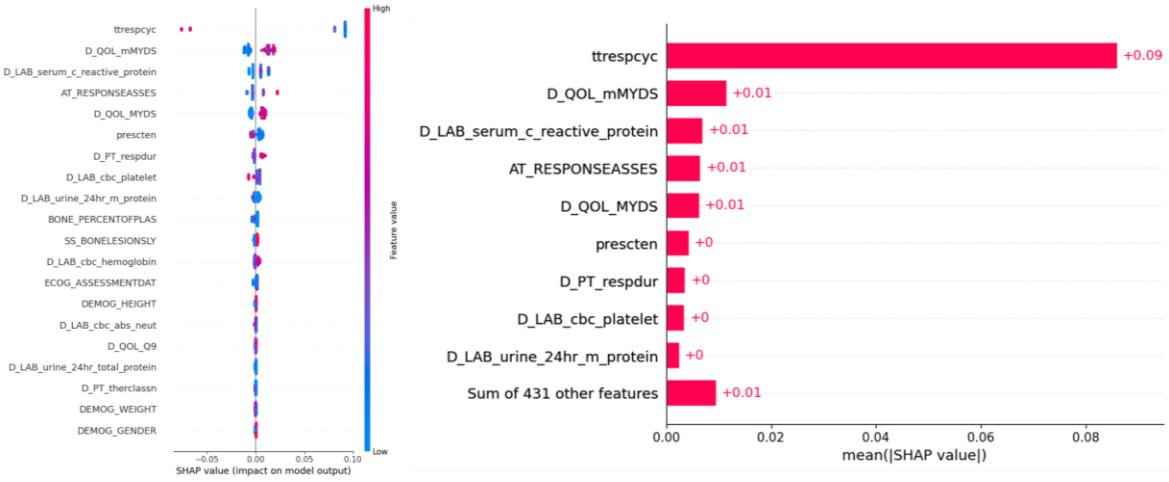

Figure 25. SHAP explainability for TR-M.1 with prediction horizon set at year 5.

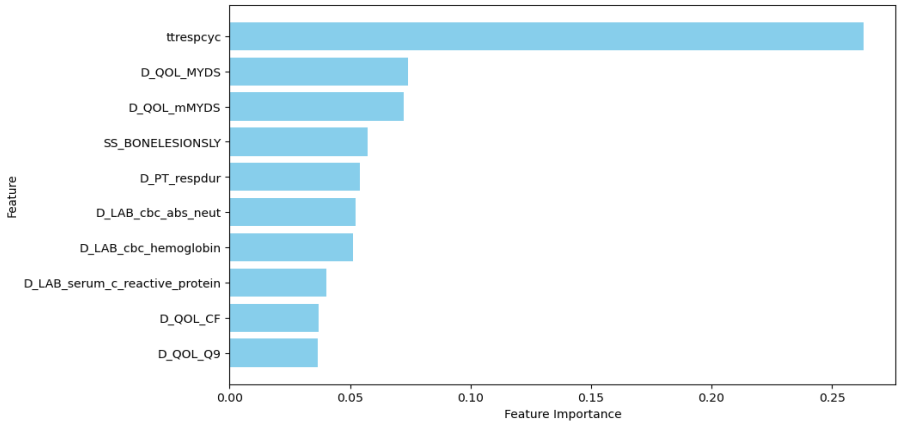

Figure 26. Feature importance (Top-10) of XGBoost model for TR-M.1 with prediction horizon at year 5.

TR-M.2. Treatment response prediction at Year 1 and Year 5 (Longitudinal approach)

In an effort to integrate temporal features and exploit the longitudinal nature of data in modeling first-line treatment response as a three-class classification problem, training results were moderate (Table 6). The best performing model is Attention LSTM model at the prediction horizon 'Year 1' for first line treatment response. An accuracy of 73.9% suggests that the model correctly classifies treatment responses in about three-quarters of the cases, which is reasonably good for a multi-class problem but leaves room for refinement. The recall of 60.8% indicates that the model is able to correctly identify 60.8% of all relevant treatment responses. This relatively moderate recall implies that the model misses a notable portion of true treatment responses, which could be critical in a medical context where recognizing all possible outcomes is important. With a precision of 69.1%, the model's predictions are correct 69.1% of the time when it assigns a treatment response class. This suggests a moderate rate of false positives, where the model predicts a certain treatment response that doesn't occur in reality. Overall, while the model demonstrates reasonable accuracy and precision, its lower recall and F1-score suggest the need for further enhancement, particularly in correctly capturing all relevant treatment responses to ensure comprehensive and reliable predictions in a clinical setting. Distribution of accuracy and loss during training and across all epochs is shown in Figure 29-Figure 30 for each prediction horizon and model. Both models' architectures are depicted in Figure 27-Figure 28.

Table 6. TR-M.2 prediction results

AI/ML Model	Prediction window	Final hyperparameters	Imbalance ratio	Performance metrics (%)
LSTM CNN	Year 1	Batch size: 8 Learning rate: 1e-05, Number of epochs: 50 Optimizer: RMSProp LSTM units: 64 CNN filters: 64	Partial Response samples:162 Complete response samples: 39 No response samples: 16	**Accuracy**: 69.6 **Recall:** 49.7 **Precision:**52.5 **F1-score:** 50.1

		Dense units: 32 Dropout rate: 0.1		
Attention LSTM	Year 1	Batch size: 8 Learning rate: 1e-04, Number of epochs: 50 Optimizer: RMSProp LSTM units: 16 Dense units: 32	Partial Response samples: 162 Complete response samples: 39 No response samples: 16	**Accuracy**: 73.9 **Recall:** 60.8 **Precision:** 69.1 **F1-score:** 63.6
LSTM CNN	Year 5	Batch size: 4 Learning rate: 1e-04, Number of epochs: 50 Optimizer: RMSProp LSTM units: 32 CNN filters: 16 Dense units: 64 Dropout rate: 0.3	Partial Response samples: 96 Complete response samples: 85 No response samples: 36	**Accuracy**: 74.0 **Recall:** 42.5 **Precision:** 58.7 **F1-score:** 44.7
Attention LSTM	Year 5	Batch size: 16 Learning rate: 1e-04, Number of epochs: 50 Optimizer: Adam LSTM units: 16 Dense units:128	Partial Response samples: 96 Complete response samples: 85 No response samples: 36	**Accuracy**: 69.6 **Recall:** 58.8 **Precision:** 71.5 **F1-score:** 57.7

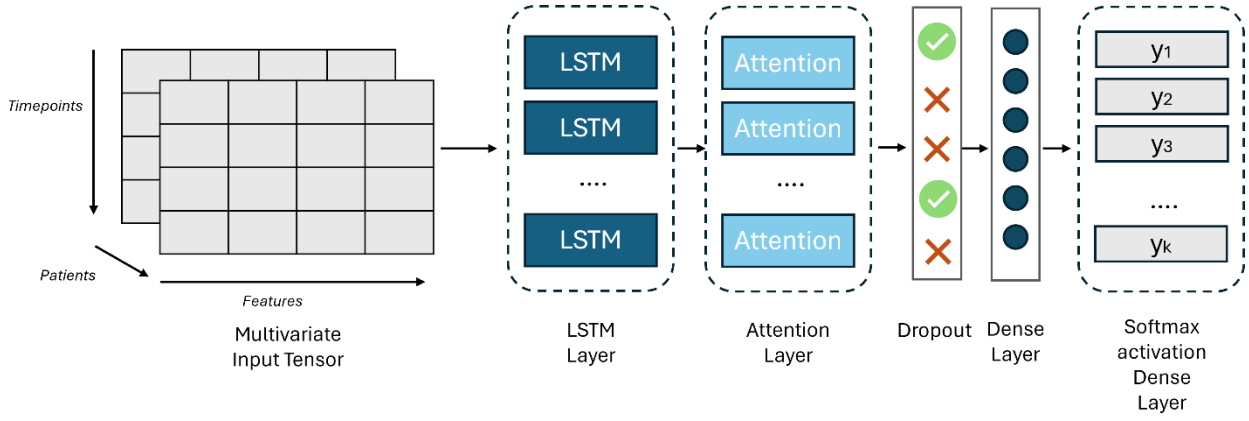

Figure 27. Attention-LSTM architecture for TR-M.2 prediction.

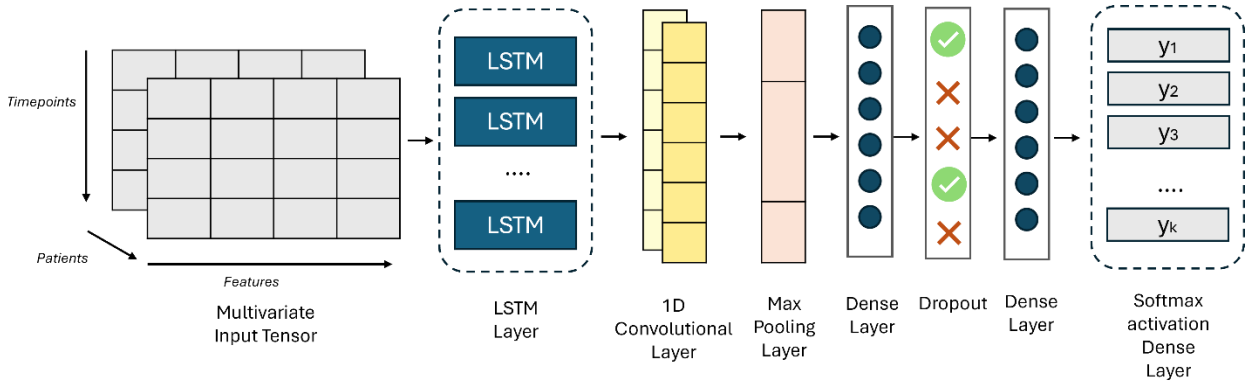

Figure 28. Figure 27. LSTM-CNN architecture for TR-M.2 prediction.

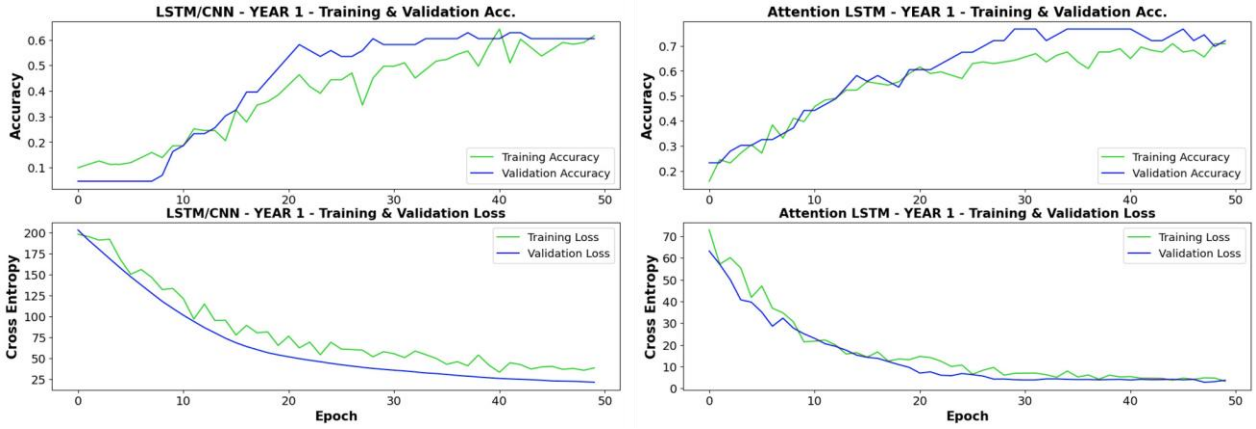

Figure 29. (Left-top) Training and validation accuracy curves of LSTM-CNN model, (Left-down) Training and validation loss curves of LSTM-CNN model, (Right-top) Training and validation accuracy curves of Attention-LSTM model, (Right-down) Training and validation loss curves of Attention-LSTM model. Treatment response prediction horizon is set at one year.

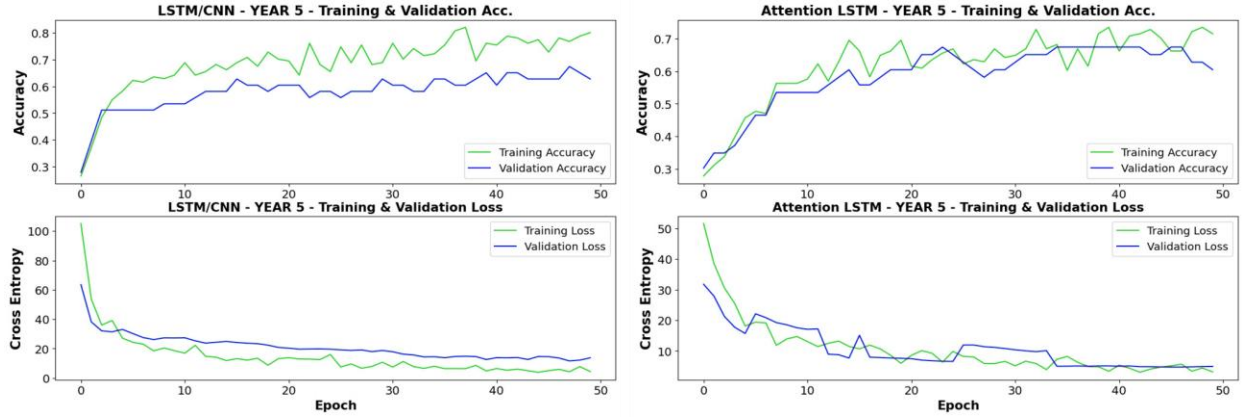

Figure 30. (Left-top) Training and validation accuracy curves of LSTM-CNN model, (Left-down) Training and validation loss curves of LSTM-CNN model, (Right-top) Training and validation accuracy curves of Attention-LSTM model, (Right-down) Training and validation loss curves of Attention-LSTM model. Treatment response prediction horizon is set at five years.

TR-M.3. Treatment response prediction every 6 months ahead (Sliding windows Approach)

This approach of exploiting temporal dependencies for predicting treatment response binary label every six months, results in high performing DL models (Table 7). Overall, the Attention-LSTM outperforms the other two models, with great discriminative ability as well as high sensitivity. The final architecture of the models is depicted in Figure 31-Figure 33-Figure 35. The training and validation curves (Figure 32-Figure 34-Figure 36) show the models have a tendency to overfit, but balance has been kept achieving both high accuracy and low loss values.

Table 7. TR-M.3 prediction results

AI/ML Model	Prediction window	Final hyperparameters	Imbalance ratio	Performance metrics (%)
Attention LSTM	6 months	Batch size: 16 Learning rate: 1e-04 Number of epochs: 50 Optimizer: Adam LSTM units: 16 Dense units: 16 Dropout rate: 0.1	Responsive window samples: 11201 Not responsive window samples: 3718	**Accuracy**: 80.7 **Recall:** 83.2 **Precision:**83.5 **F1-score:** 91.2
Attention Bidirectional LSTM	6 months	Batch size: 32 Learning rate: 1e-04 Number of epochs: 120 Optimizer: SGD LSTM units: 32 Dense units: 64 Dropout rate: 0.5	Responsive window samples: 11201 Not responsive window samples: 3718	**Accuracy**: 78.3 **Recall:** 81.2 **Precision:**83.6 **F1-score:** 93.2
LSTM-CNN	6 months	Batch size: 8 Learning rate: 1e-04 Number of epochs: 20 Optimizer: Adam LSTM units: 16 CNN filters: 16 Dense units: 16	Responsive window samples: 11201 Not responsive window samples: 3718	**Accuracy**: 79.0 **Recall:** 83.7 **Precision:**83.6 **F1-score:** 87.8

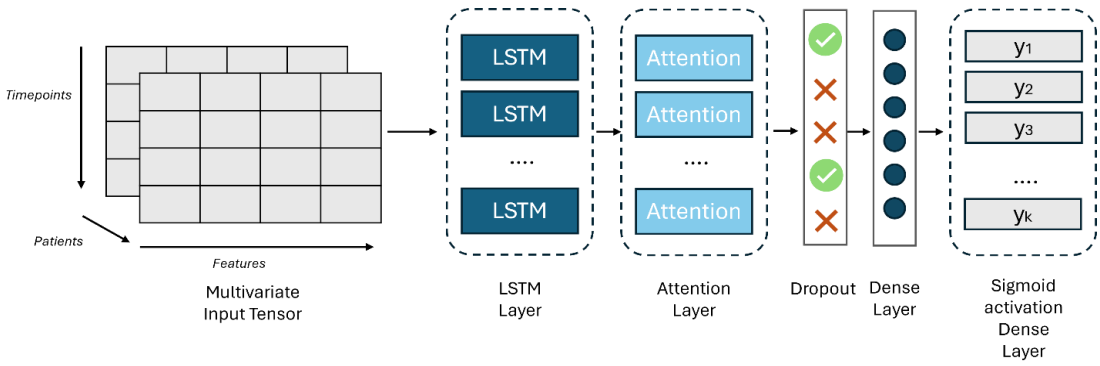

Figure 31. Attention-LSTM architecture for TR-M.3 prediction.

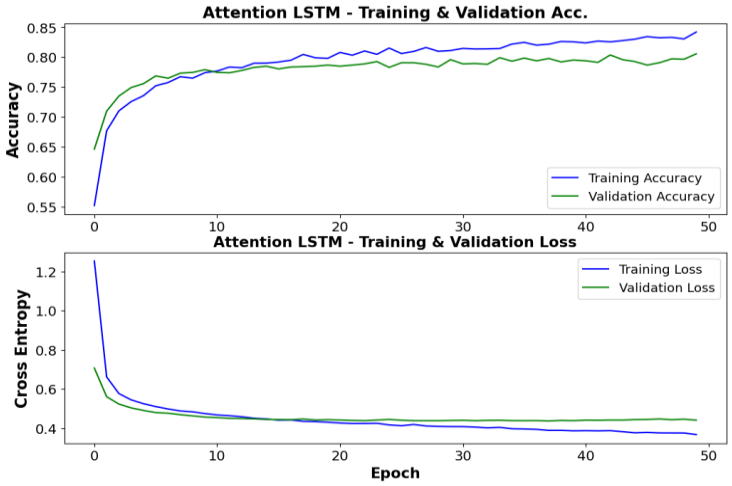

Figure 32. (Top) Training and validation accuracy curves, (Down) Training and validation loss curves for Attention-LSTM model

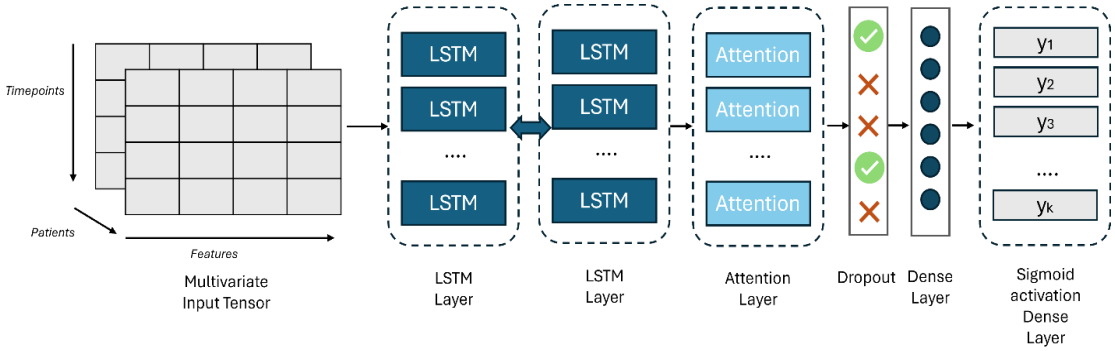

Figure 33. Attention Bi-LSTM model architecture for TR-M.3 prediction.

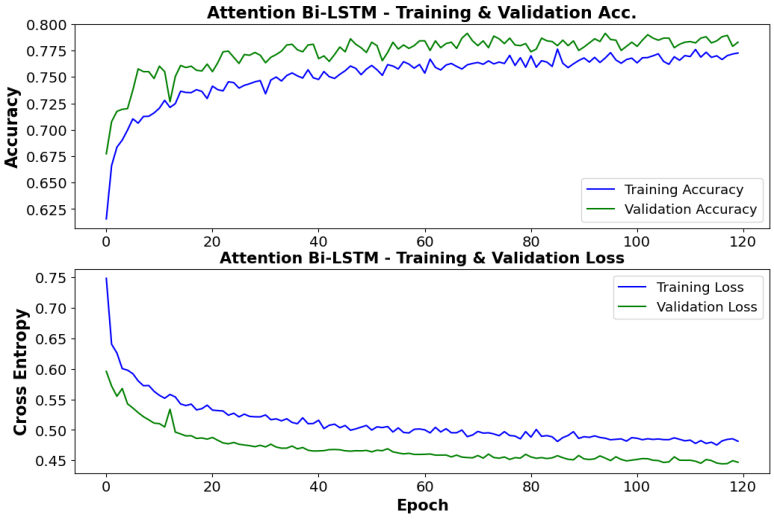

Figure 34. (Top) Training and validation accuracy curves, (Down) Training and validation loss curves for Attention Bi-LSTM model

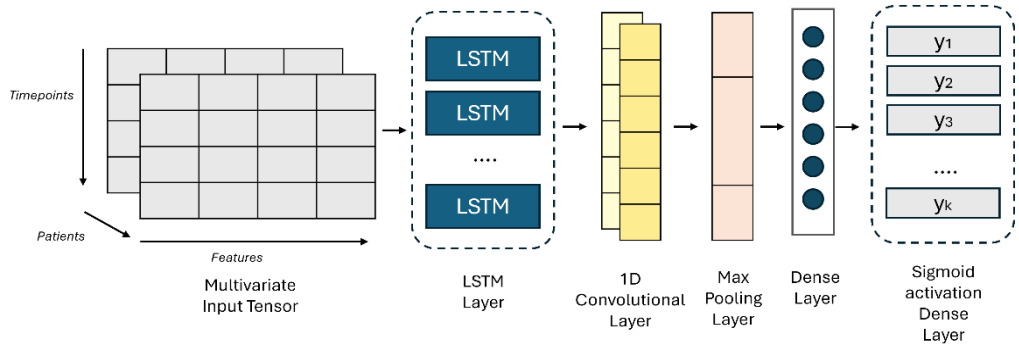

Figure 35. LSTM-CNN model architecture for TR-M.3 prediction.

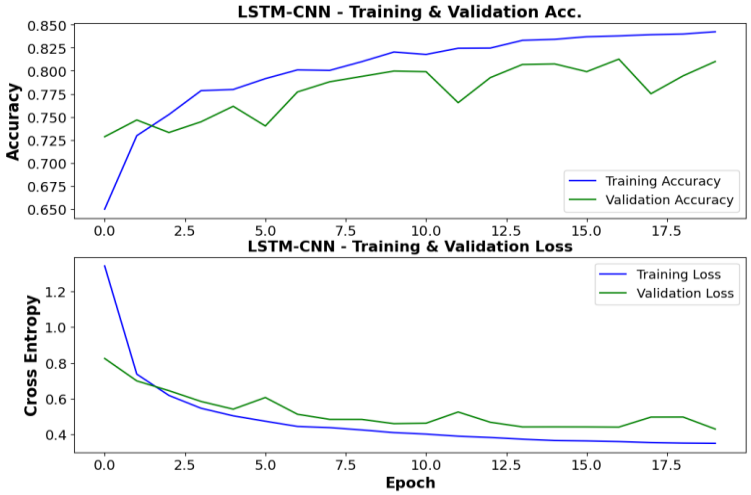

Figure 36. (Top) Training and validation accuracy curves, (Down) Training and validation loss curves for LSTM-CNN model

Mortality Risk prediction

MR-M.1. Mortality prediction within 5-years since baseline

The interpretation of the results from the XGBoost model predicting mortality underscores its commendable performance across key evaluation metrics (Table 8). With a respectable accuracy score, the model showcases its ability to effectively classify instances into their respective mortality outcomes. A balanced recall indicates the model's capacity to identify a considerable portion of actual positive cases, demonstrating its effectiveness in capturing instances of mortality among the population. Furthermore, the high precision score signifies the model's aptitude for accurately classifying positive predictions, thus minimizing false positives. These findings collectively highlight the XGBoost model's robustness in identifying mortality risk, thus offering valuable insights for clinical decision-making and patient management strategies. Based on the model results interpretability, the feature importance plot reveals that the status of MM at baseline ('D_PT_mmstatus') is the most influential feature in predicting the outcome, which is also verified from the SHAP explainability plots. Other features such as duration of treatment response

('D_PT_respdur') and lab measuremens such as plasma cells in bones (%), Albumin (g/L), Total Protein (g/dL) and Serum Kappa (mg/dL) are also found to have a significant impact on model's predictions.

Table 8. MR-M.1 prediction results

AI/ML Model	Prediction window	Final hyperparameters	Imbalance ratio	Performance metrics (%)
XGBoost	5 years	Number of estimators: 200 Learning rate: 0.1 Subsampling factor: 0.8 Column sampling factor: 0 Maximum depth: 3 Minimum child weight: 0 lambda parameter: 1.0 alpha parameter: 0	Negative samples: 636 Positive samples: 275	**Accuracy**: 75.6 **Recall:** 64.1 **Precision:** 81.9 **F1-score:** 64.5

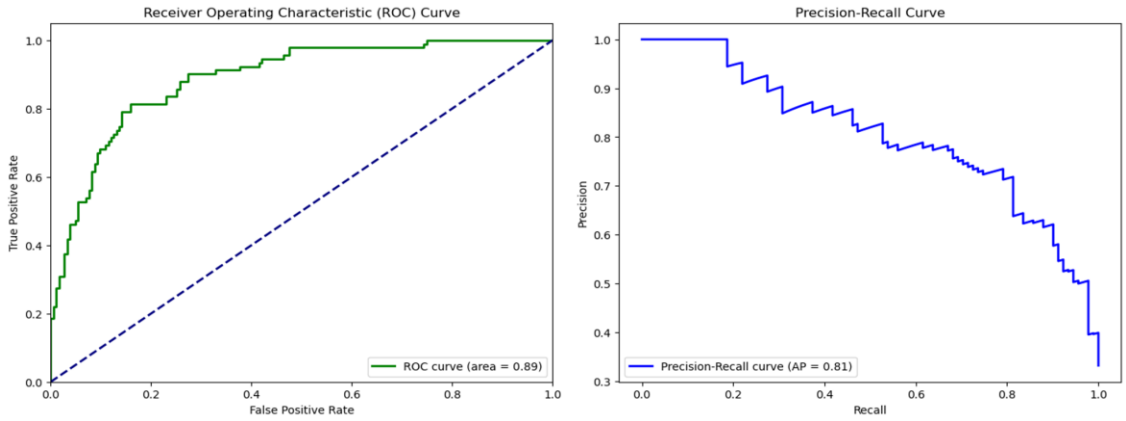

Figure 37. (Left) ROC curve and (Right) Precision-Recall curve for MR-M.1 prediction.

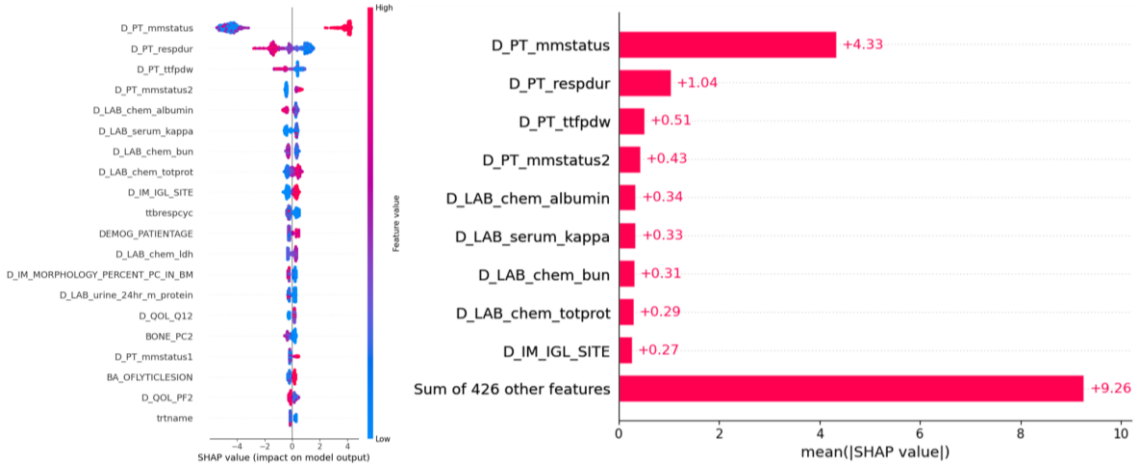

Figure 38. SHAP explainability for MR-M.1 prediction.

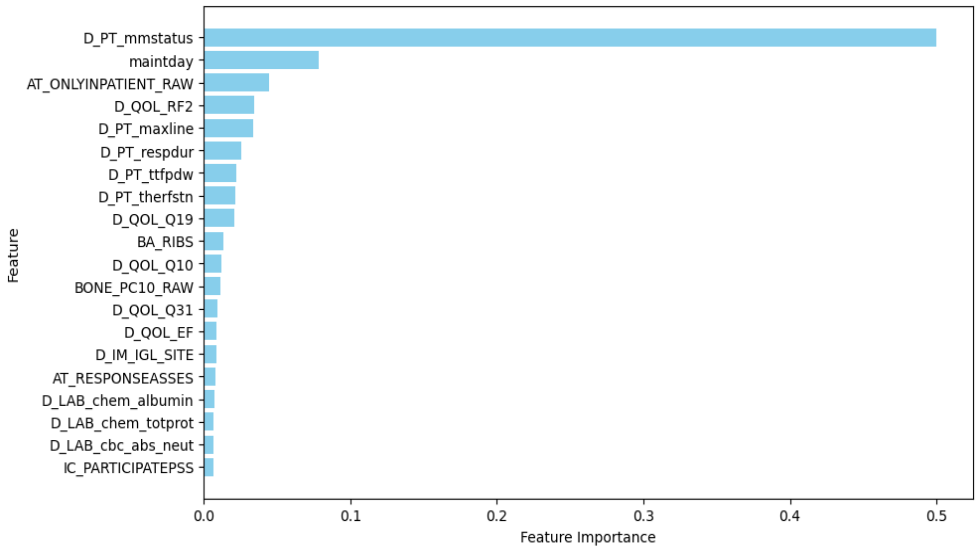

Figure 39. Feature importance (Top-20) of XGBoost model for MR-M.1 prediction.

MR-M.2. Mortality prediction within 5-years for patient clusters

For the second approach, where K-Means is implemented at the first stage, the final number of clusters is 2 (Figure 40). The maximal number of clusters examined to represent the data was 10, but based on the Silhouette analysis and Elbow Method, the optimal number of clusters selected was 2. The total number of PCA components was determined utilizing the Variance Explained and Scree plot, selecting enough principal components to account for 95% of the total variance in the dataset, which resulted in selecting 563 components. The visualization of the first two principal components of the clusters formed by K-Means clustering in the reduced-dimensional space obtained through PCA, is depicted in Figure 40.

Some details of each cluster are provided in Table 9. From this analysis, it is evident that mortality rate is almost double in Cluster 2, while ISS Stage III patients prevail as the majority compared to Cluster 1, where most patients' disease is classified as ISS Stage I.

Based on the clusters created, both t-test and Chi-square statistical tests were applied to compare the two clusters' population characteristics, for continuous and categorical variables respectively. Since a lower p-value (typically ≤ 0.05) indicates stronger evidence against the null hypothesis, suggesting that there is a significant difference between the groups for that column, it was considered the main criterion along with the t-statistic and Chi-square test values, to interpret the results from the statistical analysis. Specifically, higher absolute values of the t-statistic indicate a greater difference between the means of the two groups, whereas higher values of the chi-squared statistic indicate a greater difference in the distribution of categories between the two groups. The 8 most differentiable features between the two clusters are given in Table 10-Table 11 for each statistical test.

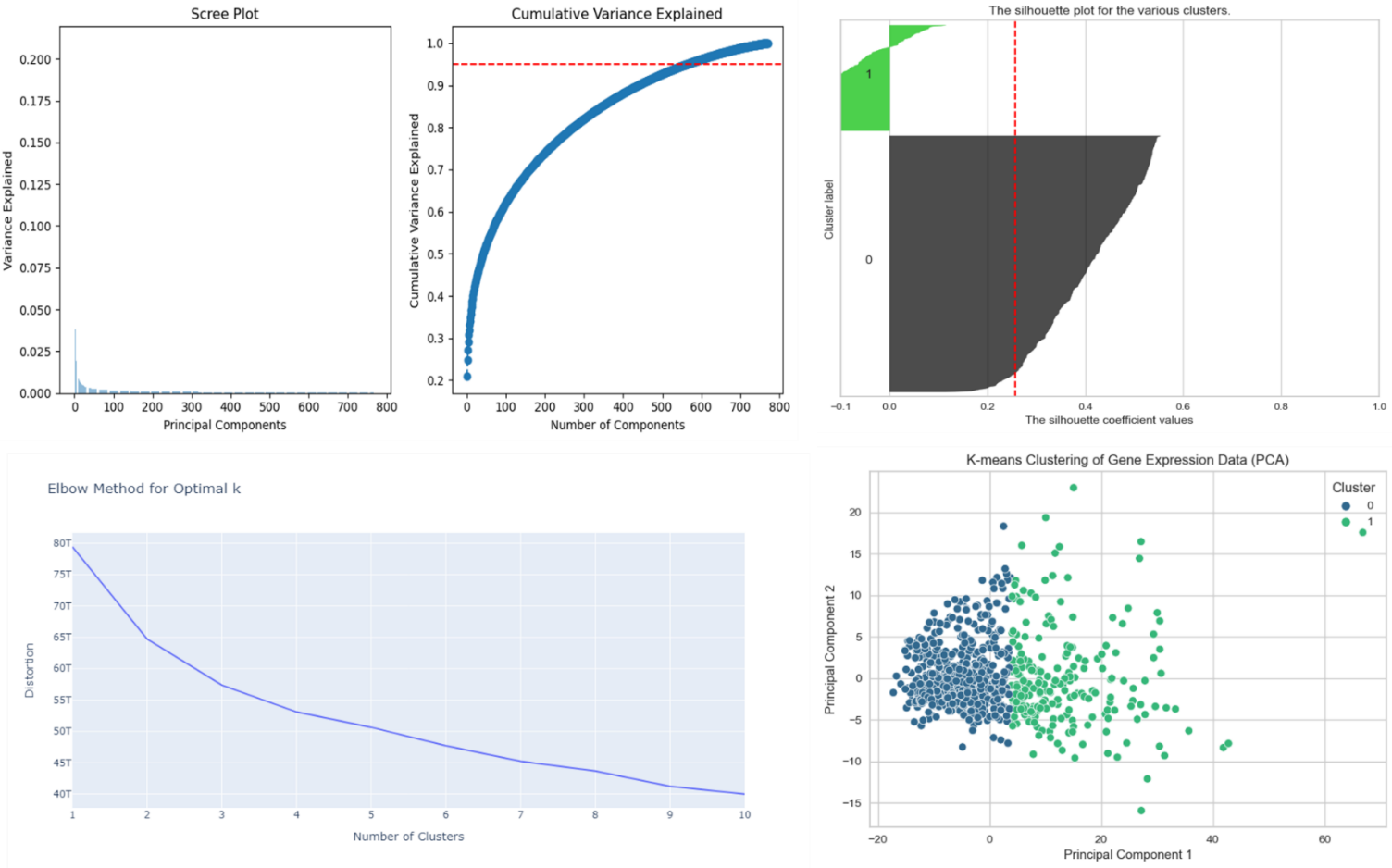

Figure 40. (Left-top) Scree plot and Cumulative Variance explained, (Right-top) Silhouette analysis plot, (Left-down) Elbow method, (Right-down) PCA analysis for two components.

Table 9. Clusters details

Cluster	Number of patients	Creatinine (µmol/L)	Age	Gender	ISS Stage I	ISS Stage II	ISS Stage III	5-year mortality
Cluster 1	426	123.994	63.68	Female: 243 Male: 183	158	154	114	21.43%
Cluster 2	176	125.924	64.09	Female:120 Male: 56	57	56	63	39.20%

Table 10. T-test results

Feature	T-statistic	P-value
Proliferation index	7.653	8.84106e-13
Morphology %PC in bone marrow	4.366	1.79635e-05

Status of MM	4.095	5.29261e-05
Definite development of new bone lesions or soft tissue plasmacytomas or definite increase in the size of existing bone lesions or soft tissue plasmacytomas	3.576	0.000414929
Maximum line of therapy	3.513	0.000516201
Platelet count x10^9/L	3.371	0.000831269
Reported as percentage of cells positive by Flow Cytometry (FLOW)	3.311	0.00105868
Duration of treatment response	3.2098	0.00145642

Table 11. Chi-squared test results

Feature	Chi-squared statistic	P-value
Reason for Bone Marrow Assessment	792.539447	2.405697e-52
Plasma cells in bones (%)	631.921922	2.398269e-28
Plasma cells in clot section or BM biopsy: (%PC positive for CD138 by IHC) - %PC high end of range	483.45	5.551962e-24
Specification of bone assessment	4978.700	1.325758e-20
B-lymphocyte antigen CD20 Plasma Cells (%)	185.05154	1.435164e-09
Side effects of treatment (from questionnaire)	1104.257	8.012745e-04
Diarrhea (from questionnaire)	23.349	5.457168e-03
Urine M protein	3.8863	4.867875e-02

The mortality prediction models were trained on each cluster's data yielding moderate performance results (Table 12). Compared with the approach of predicting 5-year mortality without the clustering step, this approach performed well on Cluster 1 data and worse on Cluster 2 data. In this case, drawbacks of clustering may have prevailed, since clustering typically divides the data into disjoint subsets based on similarities in feature space, potentially leading to fragmentation of the data and loss of valuable information. Consequently, models trained on these

fragmented subsets may lack the holistic perspective provided by the entire dataset, leading to suboptimal performance. Overall, while clustering followed by individual model training may offer insights into specific subsets of the data, its fragmentation and complexity often result in inferior performance compared to training a unified model on the entire dataset, highlighting the importance of considering the trade-offs between complexity, interpretability, and performance in model development for a specific prediction task.

Regarding predictors influence, MM status and treatment response-related variables present the most impact of prediction outcome as well as the MM status, accompanied by significant lab measurements that are significant indicators of disease state (Serum IgA, Urine M protein). In Cluster 1, feature importance further highlights the absence of Clonal Cells, high percentage of cells positive by Flow Cytometry and Urine M protein as impactful features. In Cluster 2, additional important features include renal insufficiency attributable to myeloma and high flow cytometer percentage of plasma cells in peripheral blood.

Table 12. MR-M.2 prediction results

AI/ML Model	Prediction window	Final hyperparameters	Imbalance ratio	Performance metrics (%)
XGBoost Cluster 1	5 years	Number of estimators: 500 Learning rate: 0.1 Subsampling factor: 0.6 Column sampling factor: 0.6 Maximum depth: 5 Minimum child weight: 3.0 lambda parameter: 1.0 alpha parameter: 0 Scale positive weight: 3.382	Negative samples: 317 Positive sampSles: 109	**Accuracy**: 77.0 **Recall:** 65.9 **Precision:**87.5 **F1-score:** 67.1
XGBoost Cluster 2	5 years	Number of estimators: 100 Learning rate: 0.01 Subsampling factor: 1.0 Column sampling factor: 0.6 Maximum depth: 3 Minimum child weight: 1.0 lambda parameter: 1.0 alpha parameter: 0 Scale positive weight:1.617	Negative samples: 107 Positive samples: 69	**Accuracy**: 64.1 **Recall:** 58.1 **Precision:** 67.2 **F1-score:** 54.9

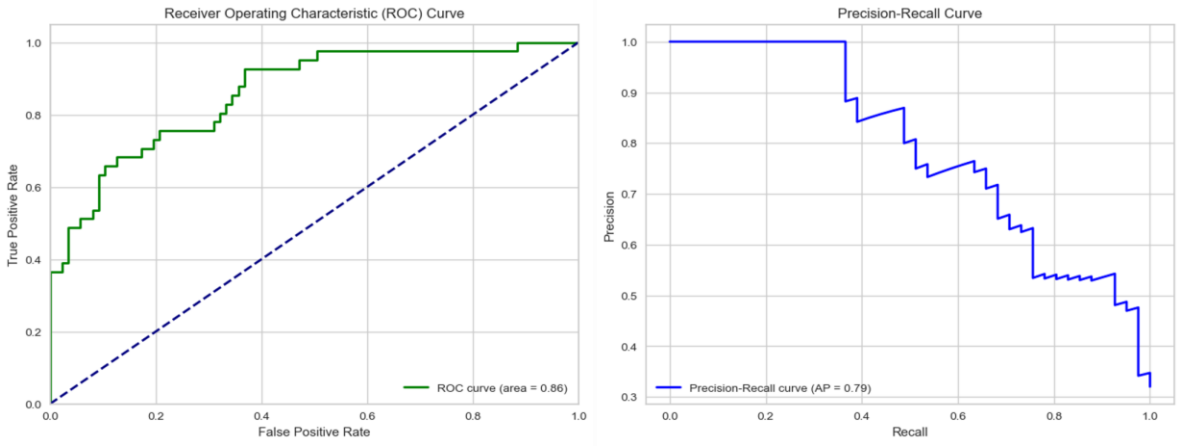

Figure 41. (Left) ROC curve and (Right) Precision-Recall curve for MR-M.2 prediction & Cluster 1.

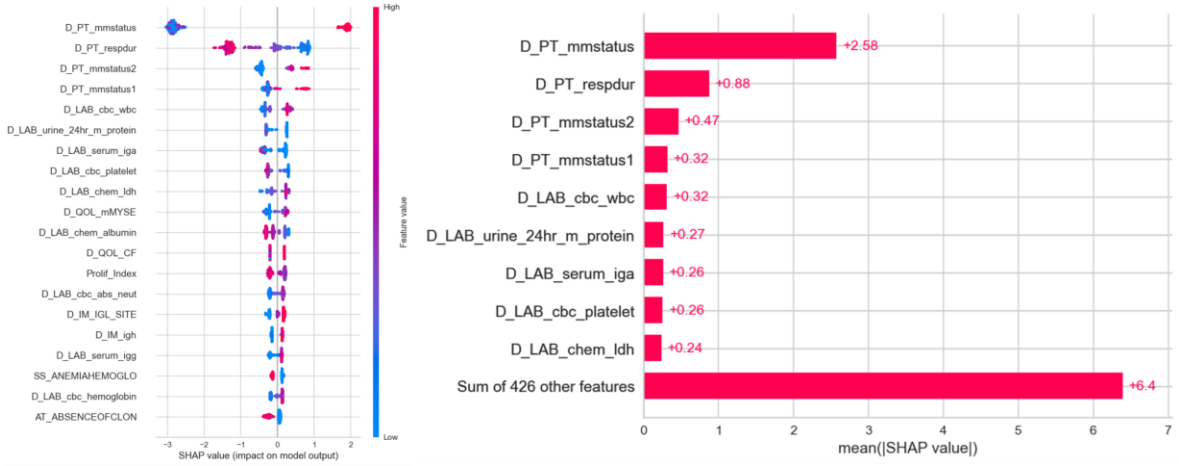

Figure 42. SHAP explainability for MR-M.2 prediction & Cluster 1.

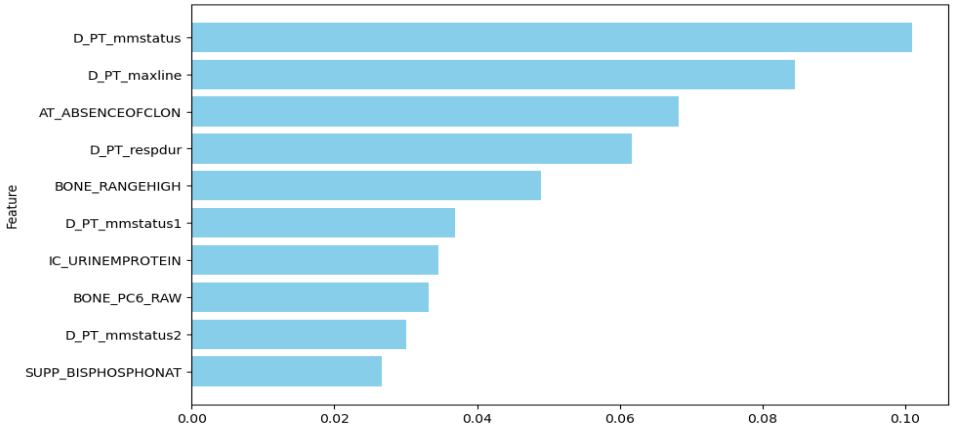

Figure 43. Feature importance (Top-10) of XGBoost model for MR-M.2 prediction & Cluster 1.

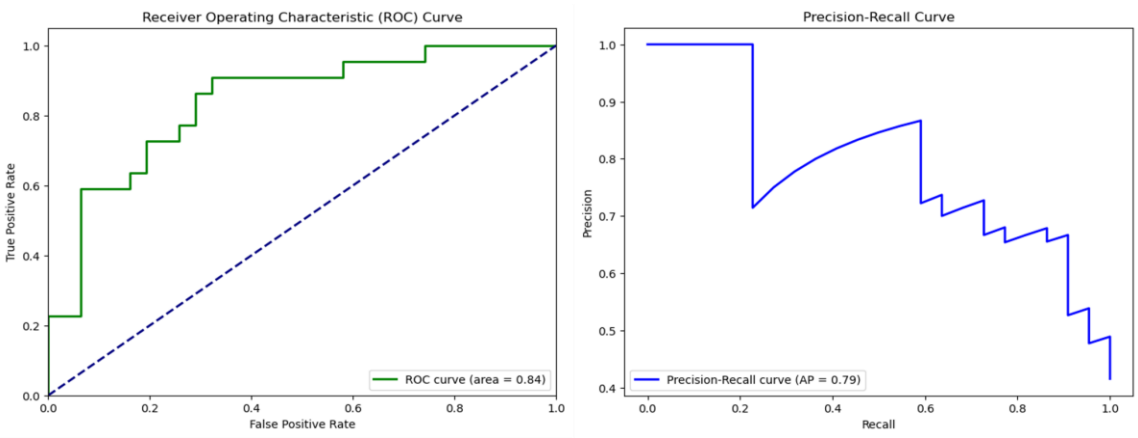

Figure 44. (Left) ROC curve and (Right) Precision-Recall curve for MR-M.2 prediction & Cluster 2.

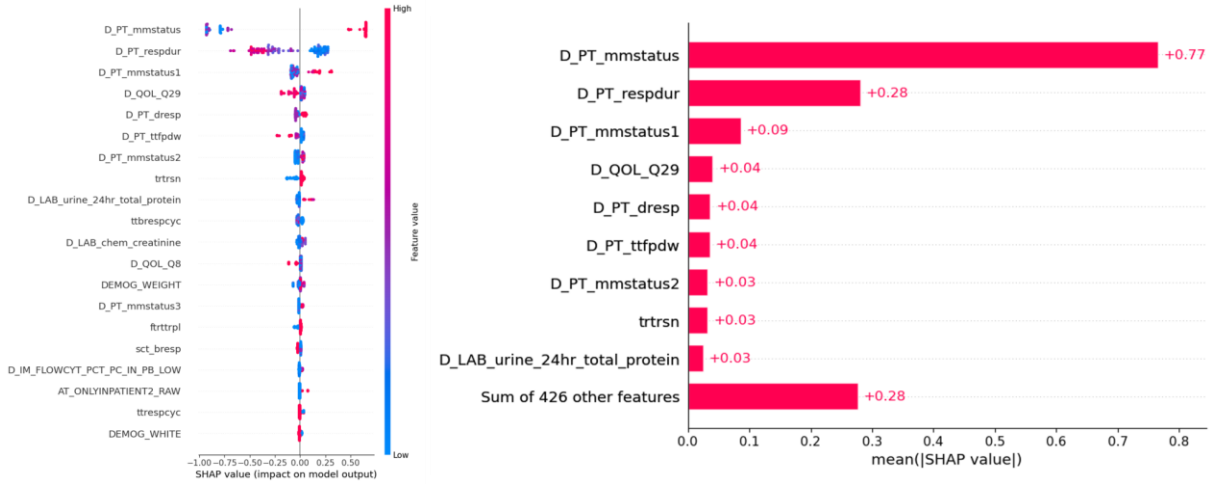

Figure 45. SHAP explainability for MR-M.2 prediction & Cluster 2.

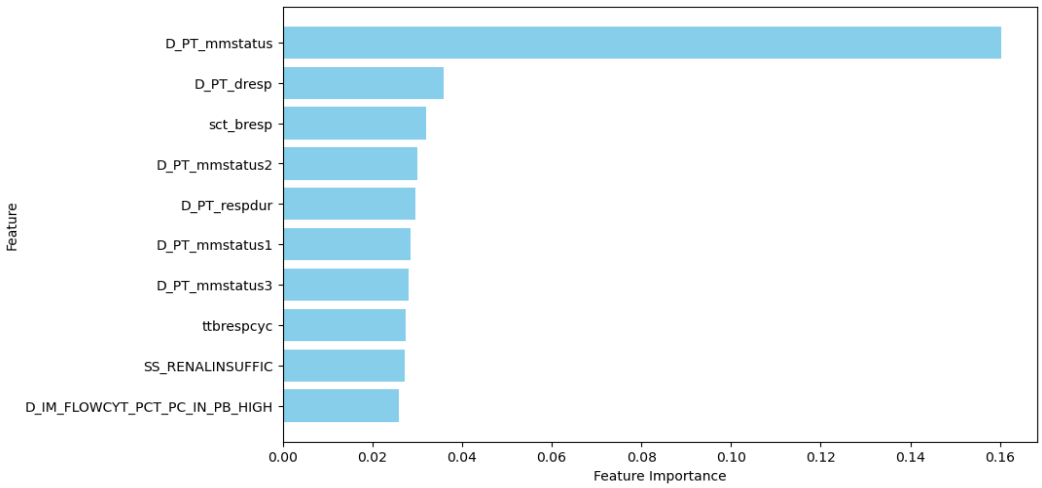

Figure 46. Feature importance (Top-10) of XGBoost model for MR-M.2 prediction & Cluster 2.

Discussion

Findings

The multifaceted nature of Multiple Myeloma, driven by a complex interaction of genetic, molecular, and clinical factors, necessitates innovative predictive modeling approaches to enhance overall patient management. This thesis aimed to leverage comprehensive multi-source data to develop accurate and reliable predictive models addressing three critical events in MM, including relapse prediction, treatment response, and mortality risk. Despite initial therapeutic responses, MM frequently relapses, signaling the emergence of drug-resistant clones or disease progression. This research focused on constructing predictive models using longitudinal clinical, phenotype and transcriptomic data to anticipate disease relapse at the first and fifth year since diagnosis and treatment initiation.

The findings suggest that the implemented models demonstrate robust predictive performance in timely identification of patients at risk for relapse. The exploitation of these forecasting models could potentially enable clinicians to implement preemptive strategies, such as salvage therapies, clinical trial enrollment, or intensified treatment regimens. These interventions could contribute to forestalling disease progression and improve quality of life, demonstrating the model's clinical utility in proactive patient management.

Predicting treatment response is crucial in MM, as it could provide guidance for therapeutic decisions, refine dosage optimization, and customize surveillance strategies according to the specific requirements of individual patients. This forecasting task was divided into three subtasks, each reflecting a different context of treatment response. Since first line treatment is clinically outlined as a significant factor of the disease status and complexity, responsiveness was predicted at first year and fifth year of first line treatment, employing both ML & DL models. Training these models as binary classifiers, input data included in one subtask only cross-sectional data collected at baseline, and in the second subtask panel data collected at each visit. These two different approaches aimed at leveraging the temporal dependencies and sequential patterns of treatment response in longitudinal data, along with reflecting a more real-world scenario where data collected at the first patient visit are used to determine patient handling, such as treatment schema selection. After training and validation of these models, both subtasks extracted well-performing models, with XGBoost model triumphing in utilizing only baseline data, and Attention-LSTM model standing out in effectively handling the sequential data. Moreover, in an effort of gaining more real-time insights about the patient’s treatment reaction, a third subtask was defined to predict response to any line therapy 6 months ahead in time, by also taking 6 months of patient history for training. As a result, the LSTM-variant models trained for this purpose yielded high performance results, indicating perhaps strong temporal patterns within the data related to the outcome, facilitating a short-term prediction of treatment response. In addition, significant predictors were highlighted for treatment response, including features related to treatment such as duration and cycle and some laboratory measurements such as C-reactive protein and Urine M protein.

As a final task, this thesis integrated clinical and transcriptomic features to develop prognostic models for 5-year survival in MM, both on the entire population as well as on extracted patient clusters, further stratifying patients into distinct risk cohorts with varying survival probabilities. The models demonstrated robust predictive performance and significant potential in

informing shared decision-making, resource allocation, and tailoring palliative care interventions. The involvement of clustering transcriptomic data to extract patient clusters followed by cluster-based mortality risk prediction, provided additional granularity and personalized risk stratification, allowing for further optimization of clusters and models. SHAP analysis, revealed significant predictors related to this outcome including the status of the disease at diagnosis, the duration of the patient being responsive to a treatment schema and some laboratory values. This is clinically interpretable since the MM status plays a crucial role in disease trajectory and eventually mortality, as well as sensitivity to any treatment.

Comparative Analysis

A notable observation from this thesis is the effective exploitation of sequential data in conjunction with the use of more advanced modeling techniques compared to the literature. This stands as an innovative solution for predicting relapse, treatment response as well as mortality risk. Moreover, there is no discrimination between treatment schemas implemented, in order to train the ML & DL models on all available treatment patterns in the data. increasing the model's generalization. Furthermore, this thesis came across a difficult task of integrating clinical, phenotype and RNA-sequencing data into the training procedure, in an attempt to build enhanced patient profiles within the data, strengthen intra-correlation of patient data, and accelerate robustness of prediction models by capturing complex relationships and patterns that may be overlooked by single-modal approaches. To this extent, combining different types of data facilitates better interpretability of model predictions, as it enables a more holistic view of the data and allows for deeper insights into the factors influencing the model's decisions. Overall, these models exhibited above average performance when compared to models reported in other research studies, showcasing their robustness and potential for clinical applicability in predicting mortality in patients with Multiple Myeloma.

Future Research

Moving forward, future research endeavors in this domain could explore several avenues to enhance the predictive capabilities and applicability of AI models in predicting critical events in patients with Multiple Myeloma. Firstly, optimization of the already implemented prediction pipelines could enhance model performances and interpretability. In addition, investigating the integration of additional data modalities, such as genetic mutations, proteomic profiles, or imaging data, could provide a more comprehensive understanding of disease progression and treatment response, thereby improving model performance. Furthermore, exploring novel feature engineering techniques or employing other advanced DL architectures may uncover latent patterns within complex multimodal data, leading to more accurate and interpretable predictive models. Lastly, focusing on interpretability and transparency in model development by reinforcing explainable artificial intelligence techniques could enhance clinicians' trust and acceptance of predictive models, fostering their integration into clinical decision-making processes.

By addressing these research directions, future studies have the potential to significantly advance the field of predictive modeling in Multiple Myeloma, ultimately improving patient outcomes and informing personalized treatment strategies.

Conclusion

This thesis highlights the significant potential of predictive modeling in improving the management of Multiple Myeloma. By leveraging comprehensive and enhanced patient data, predictive models were developed and validated for predicting key clinical events: relapse, treatment response, and mortality risk. The relapse prediction models demonstrated the capability to anticipate disease progression, enabling timely and proactive interventions. Treatment response models could effectively guide therapeutic decisions and personalized treatment plans, enhancing efficacy and minimizing adverse effects. Mortality risk models provided critical insights for stratifying patients and optimizing care pathways.

Overall, this study underscores the transformative potential of integrating advanced AI and machine learning techniques in MM care, adding significantly to the existing body of literature and clinical applicability. Continued efforts to refine these models and address the existing challenges will be crucial in realizing their full potential, ultimately contributing to better patient outcomes and more effective MM management strategies.

References

[1] "Global Cancer Observatory." Accessed: May 30, 2024. [Online]. Available: https://gco.iarc.fr/en

[2] "Cancer Today." Accessed: May 30, 2024. [Online]. Available: https://gco.iarc.who.int/today/

[3] "Cancer Tomorrow." Accessed: May 30, 2024. [Online]. Available: https://gco.iarc.who.int/today/

[4] N. Settouti and M. Saidi, "Preliminary analysis of explainable machine learning methods for multiple myeloma chemotherapy treatment recognition," *Evol. Intell.*, vol. 17, no. 1, pp. 513–533, Feb. 2024, doi: 10.1007/s12065-023-00833-3.

[5] S. V. Rajkumar *et al.*, "Lenalidomide plus high-dose dexamethasone versus lenalidomide plus low-dose dexamethasone as initial therapy for newly diagnosed multiple myeloma: an open-label randomised controlled trial," *Lancet Oncol.*, vol. 11, no. 1, pp. 29–37, Jan. 2010, doi: 10.1016/S1470-2045(09)70284-0.

[6] A. Palumbo *et al.*, "Revised international staging system for multiple myeloma: A report from international myeloma working group," *J. Clin. Oncol.*, vol. 33, no. 26, pp. 2863–2869, 2015, doi: 10.1200/JCO.2015.61.2267.

[7] S. Kumar *et al.*, "International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma," *Lancet Oncol.*, vol. 17, no. 8, pp. e328–e346, Aug. 2016, doi: 10.1016/S1470-2045(16)30206-6.

[8] T. M. Nguyen *et al.*, "Deep Learning for Human Disease Detection, Subtype Classification, and Treatment Response Prediction Using Epigenomic Data," *Biomedicines*, vol. 9, no. 11, Art. no. 11, Nov. 2021, doi: 10.3390/biomedicines9111733.

[9] G. Adam, L. Rampášek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, "Machine learning approaches to drug response prediction: challenges and recent progress," *Npj Precis. Oncol.*, vol. 4, no. 1, pp. 1–10, Jun. 2020, doi: 10.1038/s41698-020-0122-1.

[10] M. Patel *et al.*, “Machine learning-based radiomic evaluation of treatment response prediction in glioblastoma,” *Clin. Radiol.*, vol. 76, no. 8, p. 628.e17-628.e27, Aug. 2021, doi: 10.1016/j.crad.2021.03.019.
[11] L. Squarcina, F. M. Villa, M. Nobile, E. Grisan, and P. Brambilla, “Deep learning for the prediction of treatment response in depression,” *J. Affect. Disord.*, vol. 281, pp. 618–622, Feb. 2021, doi: 10.1016/j.jad.2020.11.104.
[12] A. M. Chekroud *et al.*, “The promise of machine learning in predicting treatment outcomes in psychiatry,” *World Psychiatry*, vol. 20, no. 2, pp. 154–170, Jun. 2021, doi: 10.1002/wps.20882.
[13] A. Partin *et al.*, “Deep learning methods for drug response prediction in cancer: Predominant and emerging trends,” *Front. Med.*, vol. 10, 2023, Accessed: Feb. 28, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2023.1086097
[14] R. Ettari, M. Zappalà, S. Grasso, C. Musolino, V. Innao, and A. Allegra, “Immunoproteasome-selective and non-selective inhibitors: A promising approach for the treatment of multiple myeloma,” *Pharmacol. Ther.*, vol. 182, pp. 176–192, Feb. 2018, doi: 10.1016/j.pharmthera.2017.09.001.
[15] A. Allegra *et al.*, “Novel therapeutic strategies in multiple myeloma: role of the heat shock protein inhibitors,” *Eur. J. Haematol.*, vol. 86, no. 2, pp. 93–110, 2011, doi: 10.1111/j.1600-0609.2010.01558.x.
[16] A. Allegra *et al.*, “Monoclonal antibodies: potential new therapeutic treatment against multiple myeloma,” *Eur. J. Haematol.*, vol. 90, no. 6, pp. 441–468, 2013, doi: 10.1111/ejh.12107.
[17] S. Kumar and S. V. Rajkumar, “Many facets of bortezomib resistance/susceptibility,” *Blood*, vol. 112, no. 6, pp. 2177–2178, Sep. 2008, doi: 10.1182/blood-2008-07-167767.
[18] A. Allegra, R. Ettari, V. Innao, and A. Bitto, “Potential Role of microRNAs in inducing Drug Resistance in Patients with Multiple Myeloma,” *Cells*, vol. 10, no. 2, p. 448, Feb. 2021, doi: 10.3390/cells10020448.
[19] B. Paiva *et al.*, “Phenotypic and genomic analysis of multiple myeloma minimal residual disease tumor cells: a new model to understand chemoresistance,” *Blood*, vol. 127, no. 15, pp. 1896–1906, Apr. 2016, doi: 10.1182/blood-2015-08-665679.
[20] J. Moreaux *et al.*, “A high-risk signature for patients with multiple myeloma established from the molecular classification of human myeloma cell lines,” *Haematologica*, vol. 96, no. 4, pp. 574–582, Apr. 2011, doi: 10.3324/haematol.2010.033456.
[21] F. Zhan, B. Barlogie, G. Mulligan, J. John D. Shaughnessy, and B. Bryant, “High-risk myeloma: a gene expression–based risk-stratification model for newly diagnosed multiple myeloma treated with high-dose therapy is predictive of outcome in relapsed disease treated with single-agent bortezomib or high-dose dexamethasone,” *Blood*, vol. 111, no. 2, p. 968, Jan. 2008, doi: 10.1182/blood-2007-10-119321.
[22] A. Allegra *et al.*, “New orally active proteasome inhibitors in multiple myeloma,” *Leuk. Res.*, vol. 38, no. 1, pp. 1–9, Jan. 2014, doi: 10.1016/j.leukres.2013.10.018.
[23] A. K. Mitra *et al.*, “A gene expression signature distinguishes innate response and resistance to proteasome inhibitors in multiple myeloma,” *Blood Cancer J.*, vol. 7, no. 6, pp. e581–e581, Jun. 2017, doi: 10.1038/bcj.2017.56.
[24] N. Borisov *et al.*, “Machine Learning Applicability for Classification of PAD/VCD Chemotherapy Response Using 53 Multiple Myeloma RNA Sequencing Profiles,” *Front. Oncol.*, vol. 11, p. 652063, Apr. 2021, doi: 10.3389/fonc.2021.652063.

[25] L. V. Povoa, C. H. C. Ribeiro, and I. T. da Silva, “Machine learning predicts treatment sensitivity in multiple myeloma based on molecular and clinical information coupled with drug response,” *PLOS ONE*, vol. 16, no. 7, p. e0254596, 2021, doi: 10.1371/journal.pone.0254596.
[26] J. Ubels, P. Sonneveld, E. H. van Beers, A. Broijl, M. H. van Vliet, and J. de Ridder, “Predicting treatment benefit in multiple myeloma through simulation of alternative treatment effects,” *Nat. Commun.*, vol. 9, no. 1, p. 2943, Jul. 2018, doi: 10.1038/s41467-018-05348-5.
[27] A. Paulus *et al.*, “Computational Modelling of Multiple Myeloma Patient Genomic Signatures to Predict Treatment Outcome,” *Blood*, vol. 132, p. 1911, Nov. 2018, doi: 10.1182/blood-2018-99-119574.
[28] A. Mosquera Orgueira *et al.*, “Survival prediction and treatment optimization of multiple myeloma patients using machine-learning models based on clinical and gene expression data,” *Leukemia*, vol. 35, no. 10, pp. 2924–2935, Oct. 2021, doi: 10.1038/s41375-021-01286-2.
[29] L. Ren *et al.*, “A Machine Learning Model to Predict Survival and Therapeutic Responses in Multiple Myeloma,” *Int. J. Mol. Sci.*, vol. 24, no. 7, Art. no. 7, Jan. 2023, doi: 10.3390/ijms24076683.
[30] S.-S. Park *et al.*, “ML-based sequential analysis to assist selection between VMP and RD for newly diagnosed multiple myeloma,” *Npj Precis. Oncol.*, vol. 7, no. 1, 2023, doi: 10.1038/s41698-023-00385-w.
[31] G. Cook *et al.*, “A clinical prediction model for outcome and therapy delivery in transplant-ineligible patients with myeloma (UK Myeloma Research Alliance Risk Profile): a development and validation study,” *Lancet Haematol.*, vol. 6, no. 3, pp. e154–e166, Mar. 2019, doi: 10.1016/S2352-3026(18)30220-5.
[32] A. S. Kubasch *et al.*, “Predicting Early Relapse for Patients with Multiple Myeloma through Machine Learning,” *Blood*, vol. 138, p. 2953, Nov. 2021, doi: 10.1182/blood-2021-151195.
[33] D. Baker *et al.*, “Predicting risk of progression in relapsed multiple myeloma using traditional risk models, focal lesion assessment with PET-CT and minimal residual disease status,” *Haematologica*, vol. 106, no. 12, Art. no. 12, Aug. 2021, doi: 10.3324/haematol.2021.278779.
[34] L. J. Sandell, “Genes and gene expression,” *Clin. Orthop.*, no. 379 Suppl, pp. S9-16, Oct. 2000, doi: 10.1097/00003086-200010001-00003.
[35] I. San Segundo-Val and C. S. Sanz-Lozano, “Introduction to the Gene Expression Analysis,” in *Molecular Genetics of Asthma*, M. Isidoro García, Ed., New York, NY: Springer, 2016, pp. 29–43. doi: 10.1007/978-1-4939-3652-6_3.
[36] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression,” *Nat. Methods*, vol. 14, no. 4, pp. 417–419, Apr. 2017, doi: 10.1038/nmeth.4197.
[37] N. K. Aaronson *et al.*, “The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology,” *J. Natl. Cancer Inst.*, vol. 85, no. 5, pp. 365–376, Mar. 1993, doi: 10.1093/jnci/85.5.365.
[38] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*, 1st ed. O’Reilly Media, Inc., 2018.

[39] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer, 2013. doi: 10.1007/978-1-4614-6849-3.
[40] G. Varotto, G. Susi, L. Tassi, F. Gozzo, S. Franceschetti, and F. Panzica, "Comparison of Resampling Techniques for Imbalanced Datasets in Machine Learning: Application to Epileptogenic Zone Localization From Interictal Intracranial EEG Recordings in Patients With Focal Epilepsy," *Front. Neuroinformatics*, vol. 15, Nov. 2021, doi: 10.3389/fninf.2021.715421.
[41] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science," in *Supervised and Unsupervised Learning for Data Science*, M. W. Berry, A. Mohamed, and B. W. Yap, Eds., Cham: Springer International Publishing, 2020, pp. 3–21. doi: 10.1007/978-3-030-22475-2_1.
[42] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philos. Transact. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.
[43] H. Abbasimehr and R. Paki, "Improving time series forecasting using LSTM and attention models," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 1, pp. 673–691, Jan. 2022, doi: 10.1007/s12652-020-02761-x.
[44] D. Barić, P. Fumić, D. Horvatić, and T. Lipic, "Benchmarking Attention-Based Interpretability of Deep Learning in Multivariate Time Series Predictions," *Entropy Basel Switz.*, vol. 23, no. 2, p. 143, Jan. 2021, doi: 10.3390/e23020143.
[45] L. A. Carrasco-Ribelles *et al.*, "Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review," *J. Am. Med. Inform. Assoc.*, vol. 30, no. 12, pp. 2072–2082, Dec. 2023, doi: 10.1093/jamia/ocad168.
[46] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Sep. 2017, pp. 6000–6010.
[47] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019, doi: 10.1016/j.neucom.2019.01.078.
[48] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks," *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1903–1911, Aug. 2017, doi: 10.1145/3097983.3098088.
[49] Y. Yang, X. Zheng, and C. Ji, "Disease Prediction Model Based on BiLSTM and Attention Mechanism," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Aug. 2019, pp. 1141–1148. doi: 10.1109/BIBM47256.2019.8983378.
[50] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for Time Series Classification," *Neural Netw.*, vol. 116, pp. 237–245, Aug. 2019, doi: 10.1016/j.neunet.2019.04.014.
[51] R. J. Desai, S. V. Wang, M. Vaduganathan, T. Evers, and S. Schneeweiss, "Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes," *JAMA Netw. Open*, vol. 3, no. 1, p. e1918962, Jan. 2020, doi: 10.1001/jamanetworkopen.2019.18962.

[52] J. Zheng *et al.*, “Clinical Data based XGBoost Algorithm for infection risk prediction of patients with decompensated cirrhosis: a 10-year (2012–2021) Multicenter Retrospective Case-control study,” *BMC Gastroenterol.*, vol. 23, no. 1, p. 310, Sep. 2023, doi: 10.1186/s12876-023-02949-3.
[53] J. Wu, *Advances in K-means Clustering: A Data Mining Thinking*. in Springer Theses. Berlin, Heidelberg: Springer, 2012. doi: 10.1007/978-3-642-29807-3.
[54] L. A. Jeni, J. F. Cohn, and F. De La Torre, “Facing Imbalanced Data–Recommendations for the Use of Performance Metrics,” in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 245–251. doi: 10.1109/ACII.2013.47.
[55] E. W. Steyerberg *et al.*, “Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures,” *Epidemiology*, vol. 21, no. 1, p. 128, Jan. 2010, doi: 10.1097/EDE.0b013e3181c30fb2.
[56] D. Wilimitis and C. G. Walsh, “Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial,” *JMIR AI*, vol. 2, no. 1, p. e49023, Dec. 2023, doi: 10.2196/49023.
[57] M. A. Little *et al.*, “Using and understanding cross-validation strategies. Perspectives on Saeb et al.,” *GigaScience*, vol. 6, no. 5, p. gix020, May 2017, doi: 10.1093/gigascience/gix020.
[58] N. Burkart and M. F. Huber, “A Survey on the Explainability of Supervised Machine Learning,” *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, Jan. 2021, doi: 10.1613/jair.1.12228.
[59] S. Park and J.-S. Yang, “Interpretable deep learning LSTM model for intelligent economic decision-making,” *Knowl.-Based Syst.*, vol. 248, p. 108907, Jul. 2022, doi: 10.1016/j.knosys.2022.108907.
[60] D. Saraswat *et al.*, “Explainable AI for Healthcare 5.0: Opportunities and Challenges,” *IEEE Access*, vol. 10, pp. 84486–84517, 2022, doi: 10.1109/ACCESS.2022.3197671.
[61] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, “Survey of Explainable AI Techniques in Healthcare,” *Sensors*, vol. 23, no. 2, p. 634, Jan. 2023, doi: 10.3390/s23020634.
[62] D. Freedman, R. Pisani, and R. Purves, *Statistics: Fourth International Student Edition*. W. W. Norton & Company, 2007.